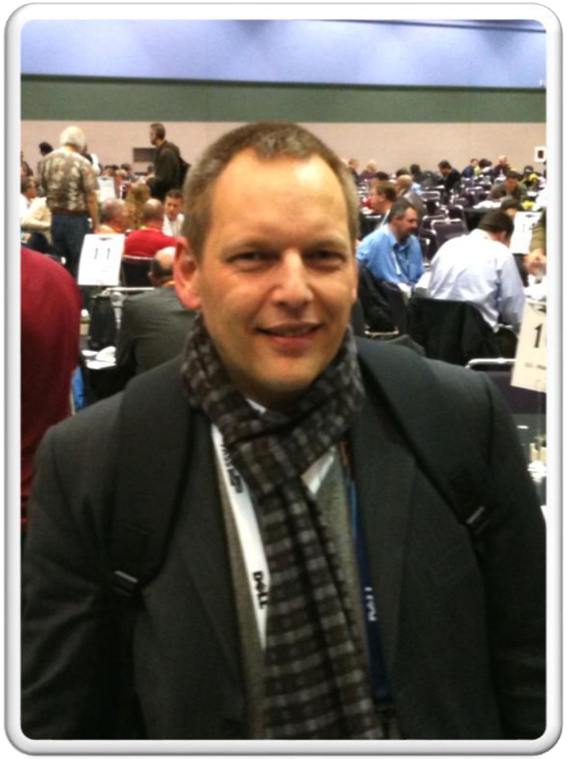
Datenqualität mit SSIS und DQS

Alexander Karl

Speaker



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

Agenda

- Definitionen rund um „Datenqualität“
- bisherige Lösungen in T-SQL und CLR
- Lösung in SSIS
- Lösung mit DQS

„Datenqualität“

- Glaubwürdigkeit
  sind die Daten in einem korrekten Intervall ?

- Interpretierbarkeit
  werden Daten fortlaufend gleich dargestellt ?

- Schlüsselintegrität
  Schlüsseleindeutigkeit / Ref. Integrität

- Nützlichkeit
  Zeitnähe, Vollständigkeit

das liefert T-SQL (1/3)

- Keine NULL Werte als Tabelleneintrag
  (default-Values)
- Constraints zw. mehreren Spalten
  (Eintrittsdatum <= Austrittsdatum)
- maximale Textlänge
  ?? *min. Textlänge* bei Nachnamen z.B. 2

das liefert T-SQL (2/3)

- Passende T-SQL Funktionen
  -- IsNumeric
  -- IsDate
  -- *IsText* ??

- Neu ab 2012
  -- Try_Cast
  -- Try_Convert
  -- Try_Parse

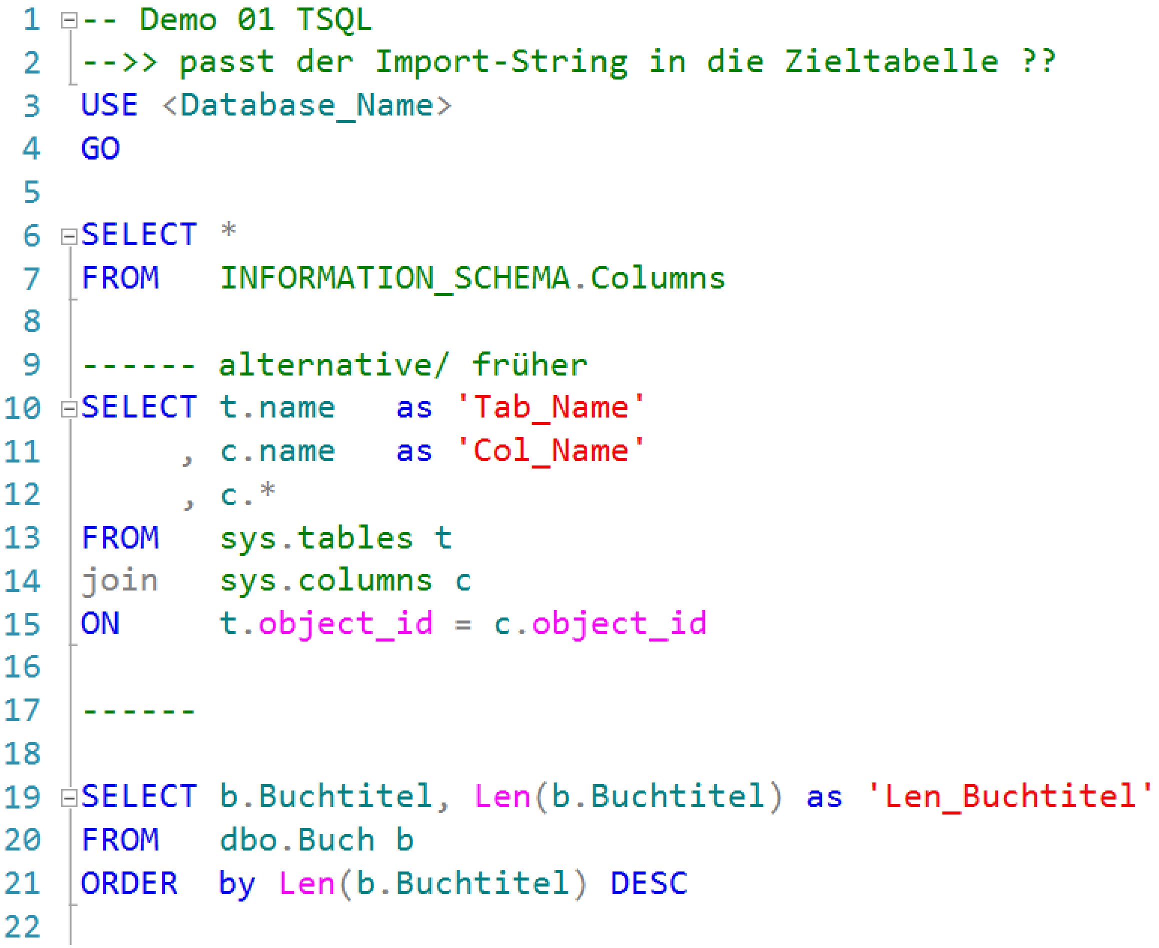
das liefert T-SQL (3/3)

- exakte Übereinstimmung
  -- WHEN IN (…)
  -- JOIN
  -- MERGE

- ähnliche Übereinstimmung
  -- Soundex (T-SQL)
  -- Difference (T-SQL)
  -- Fuzzy Lookup (SSIS Enterprise)
  -- Fuzzy Grouping (SSIS Enterprise)

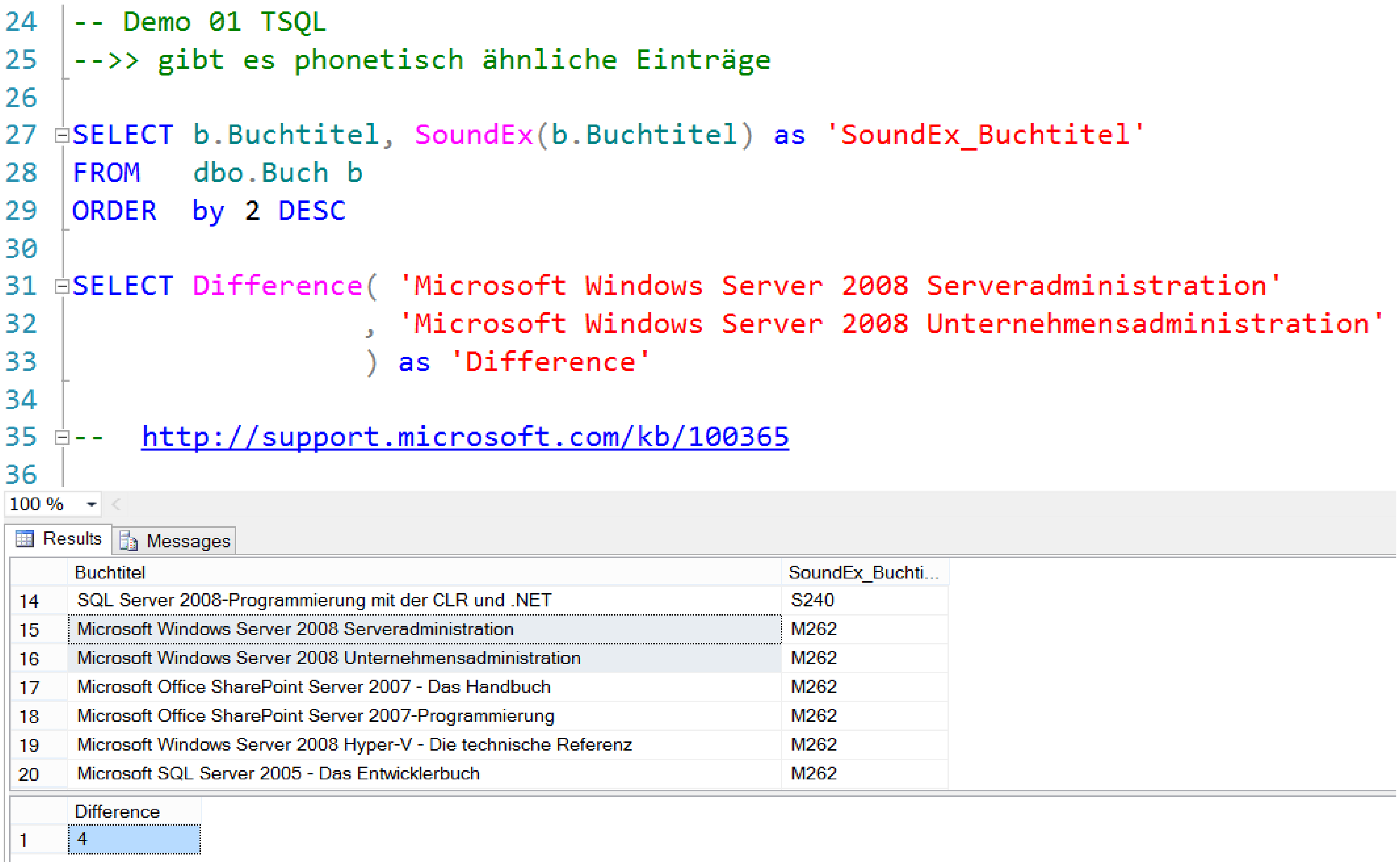
DEMO

Demo result



```sql
-- Demo 01 TSQL
-->> passt der Import-String in die Zieltabelle ??
USE <Database_Name>
GO

SELECT *
FROM   INFORMATION_SCHEMA.Columns

------ alternative/ früher
SELECT t.name   as 'Tab_Name'
     , c.name   as 'Col_Name'
     , c.*
FROM   sys.tables t
join   sys.columns c
ON     t.object_id = c.object_id

------

SELECT b.Buchtitel, Len(b.Buchtitel) as 'Len_Buchtitel'
FROM   dbo.Buch b
ORDER  by Len(b.Buchtitel) DESC

```

Demo result



```
-- Demo 01 TSQL
-->> gibt es phonetisch ähnliche Einträge

SELECT b.Buchtitel, SoundEx(b.Buchtitel) as 'SoundEx_Buchtitel'
FROM   dbo.Buch b
ORDER  by 2 DESC

SELECT Difference( 'Microsoft Windows Server 2008 Serveradministration'
                 , 'Microsoft Windows Server 2008 Unternehmensadministration'
                 ) as 'Difference'

--  http://support.microsoft.com/kb/100365

```

	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

workaround mit CLR

- CLR Functions

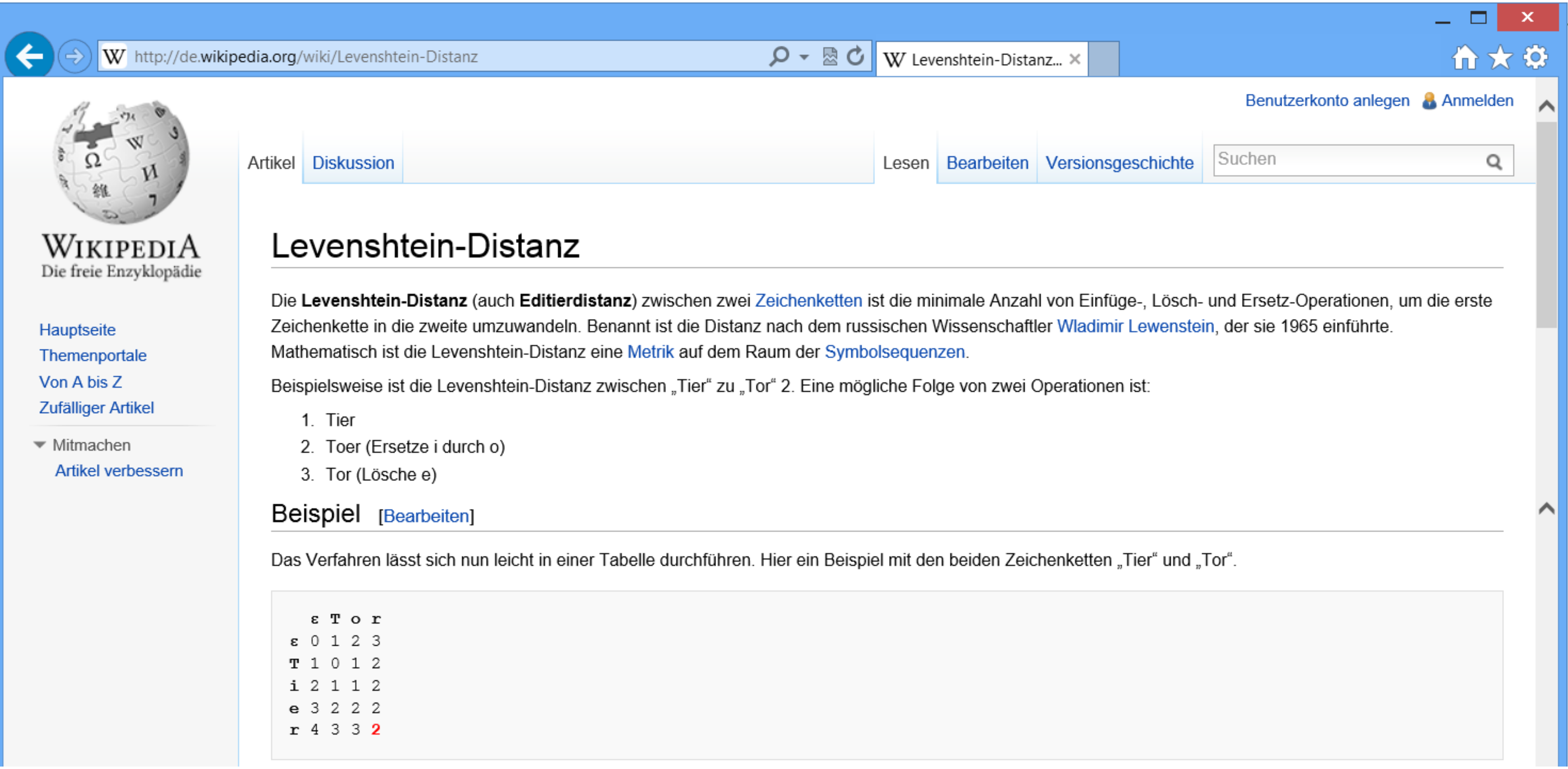
  mit RegEx
  -- Textsuche (-- *IsText*)
  mit Levenshtein
  -- ähnliche Übereinstimmung

- !! Verwendung im Dataflow

Regular Expressions

Anwendung	Regex
Buchstaben & Ziffern	\w
nur Buchstaben	^[a-zA-Z]
Zahlzeichen	\d
ISBN	^978\-\d\-\d{5}\-\d{3}\-\d
email	\b[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b

Levenshtein



Levenshtein-Distanz

Die **Levenshtein-Distanz** (auch **Editierdistanz**) zwischen zwei Zeichenketten ist die minimale Anzahl von Einfüge-, Lösch- und Ersetz-Operationen, um die erste Zeichenkette in die zweite umzuwandeln. Benannt ist die Distanz nach dem russischen Wissenschaftler Wladimir Lewenstein, der sie 1965 einführte. Mathematisch ist die Levenshtein-Distanz eine Metrik auf dem Raum der Symbolsequenzen.

Beispielsweise ist die Levenshtein-Distanz zwischen „Tier" zu „Tor" 2. Eine mögliche Folge von zwei Operationen ist:

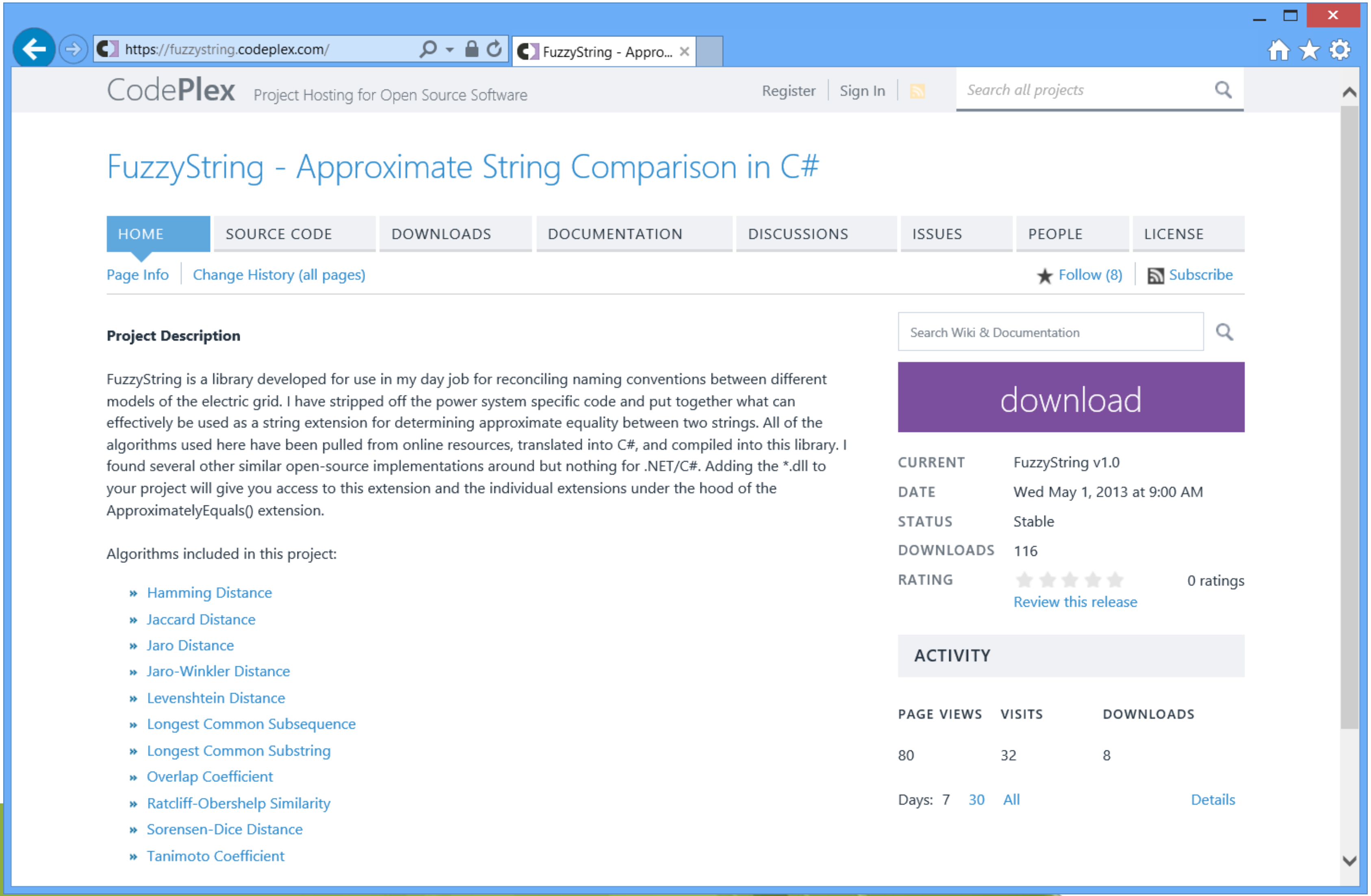
1. Tier
2. Toer (Ersetze i durch o)
3. Tor (Lösche e)

Beispiel [Bearbeiten]

Das Verfahren lässt sich nun leicht in einer Tabelle durchführen. Hier ein Beispiel mit den beiden Zeichenketten „Tier" und „Tor".

```
  ε T o r
ε 0 1 2 3
T 1 0 1 2
i 2 1 1 2
e 3 2 2 2
r 4 3 3 2
```

codeplex



https://fuzzystring.codeplex.com/
FuzzyString - Appro...
CodePlex Project Hosting for Open Source Software
Register
Sign In
Search all projects

FuzzyString - Approximate String Comparison in C#

Project Description

FuzzyString is a library developed for use in my day job for reconciling naming conventions between different models of the electric grid. I have stripped off the power system specific code and put together what can effectively be used as a string extension for determining approximate equality between two strings. All of the algorithms used here have been pulled from online resources, translated into C#, and compiled into this library. I found several other similar open-source implementations around but nothing for .NET/C#. Adding the *.dll to your project will give you access to this extension and the individual extensions under the hood of the ApproximatelyEquals() extension.

Algorithms included in this project:

- Hamming Distance
- Jaccard Distance
- Jaro Distance
- Jaro-Winkler Distance
- Levenshtein Distance
- Longest Common Subsequence
- Longest Common Substring
- Overlap Coefficient
- Ratcliff-Obershelp Similarity
- Sorensen-Dice Distance
- Tanimoto Coefficient

CURRENT	FuzzyString v1.0
DATE	Wed May 1, 2013 at 9:00 AM
STATUS	Stable
DOWNLOADS	116
RATING	☆☆☆☆☆ 0 ratings Review this release

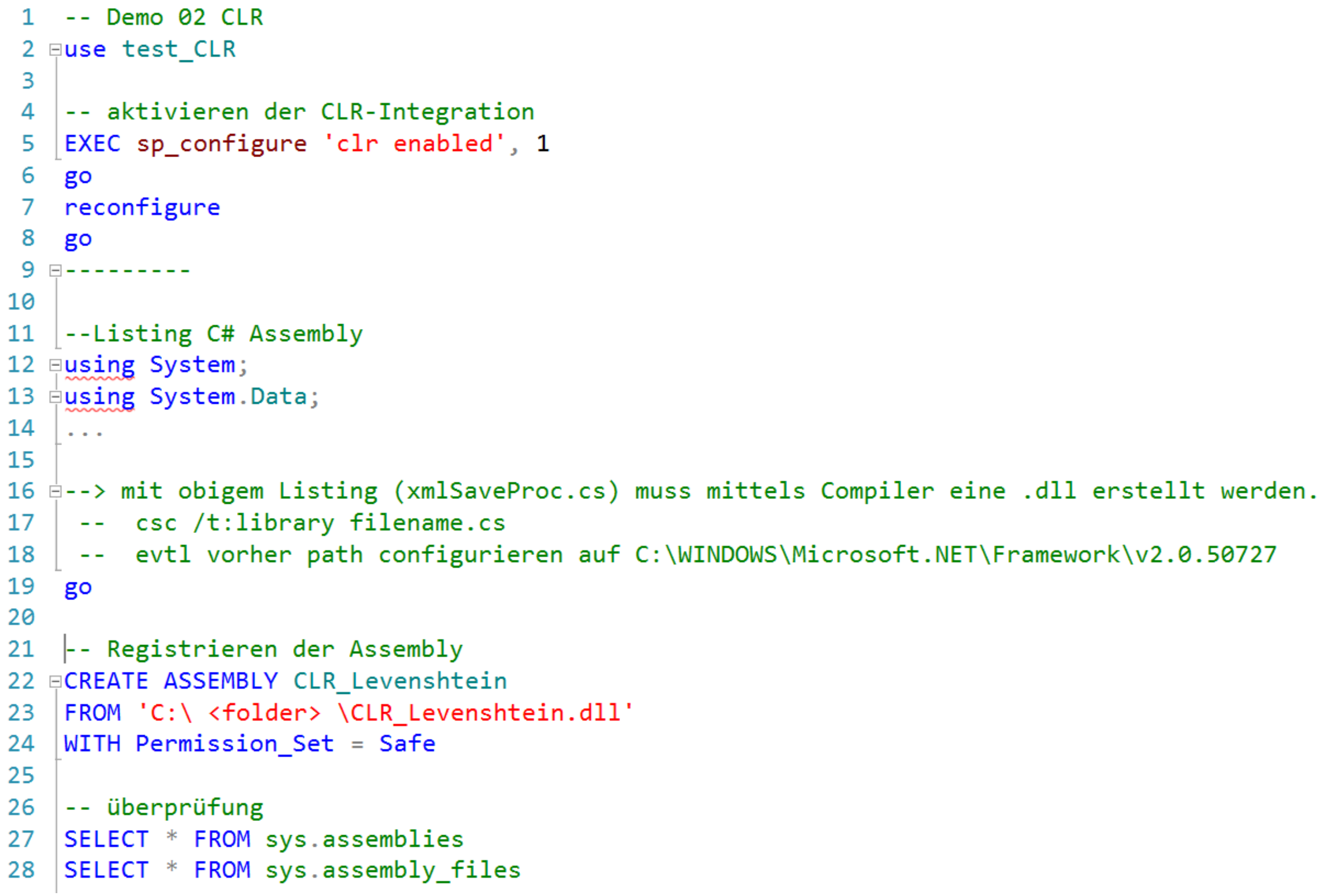
ACTIVITY

PAGE VIEWS	VISITS	DOWNLOADS
80	32	8

Days: 7 30 All Details

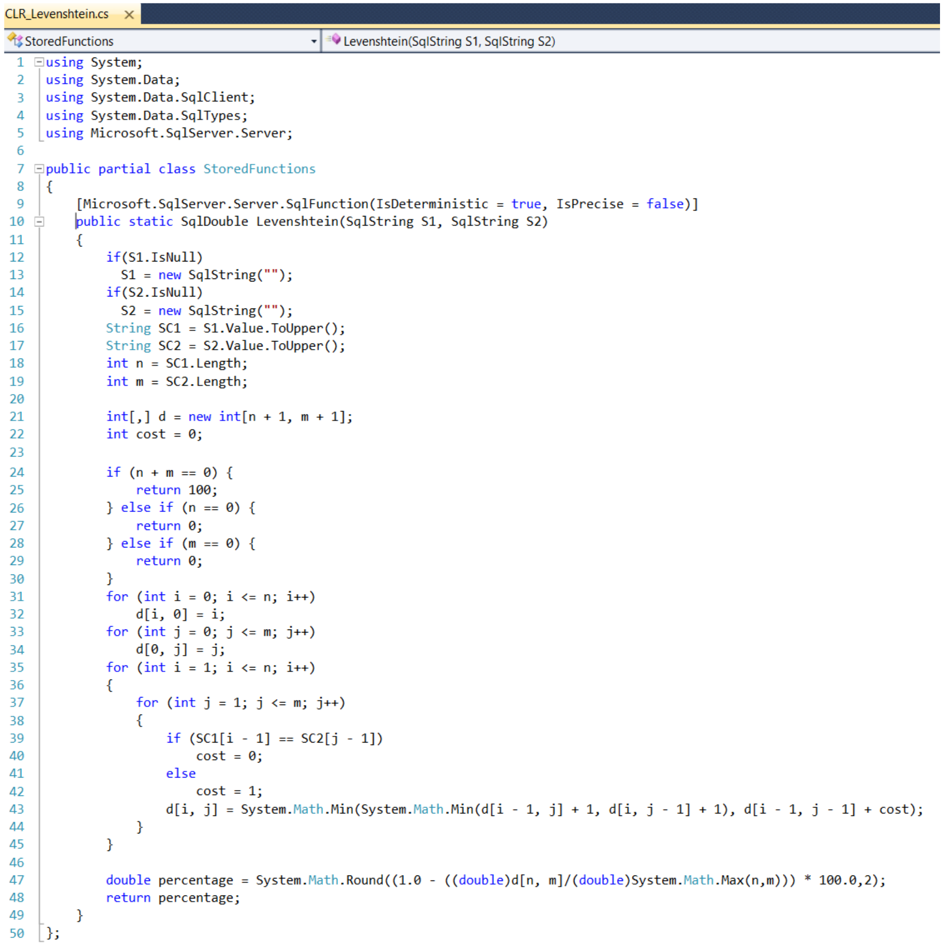
DEMO

Demo result



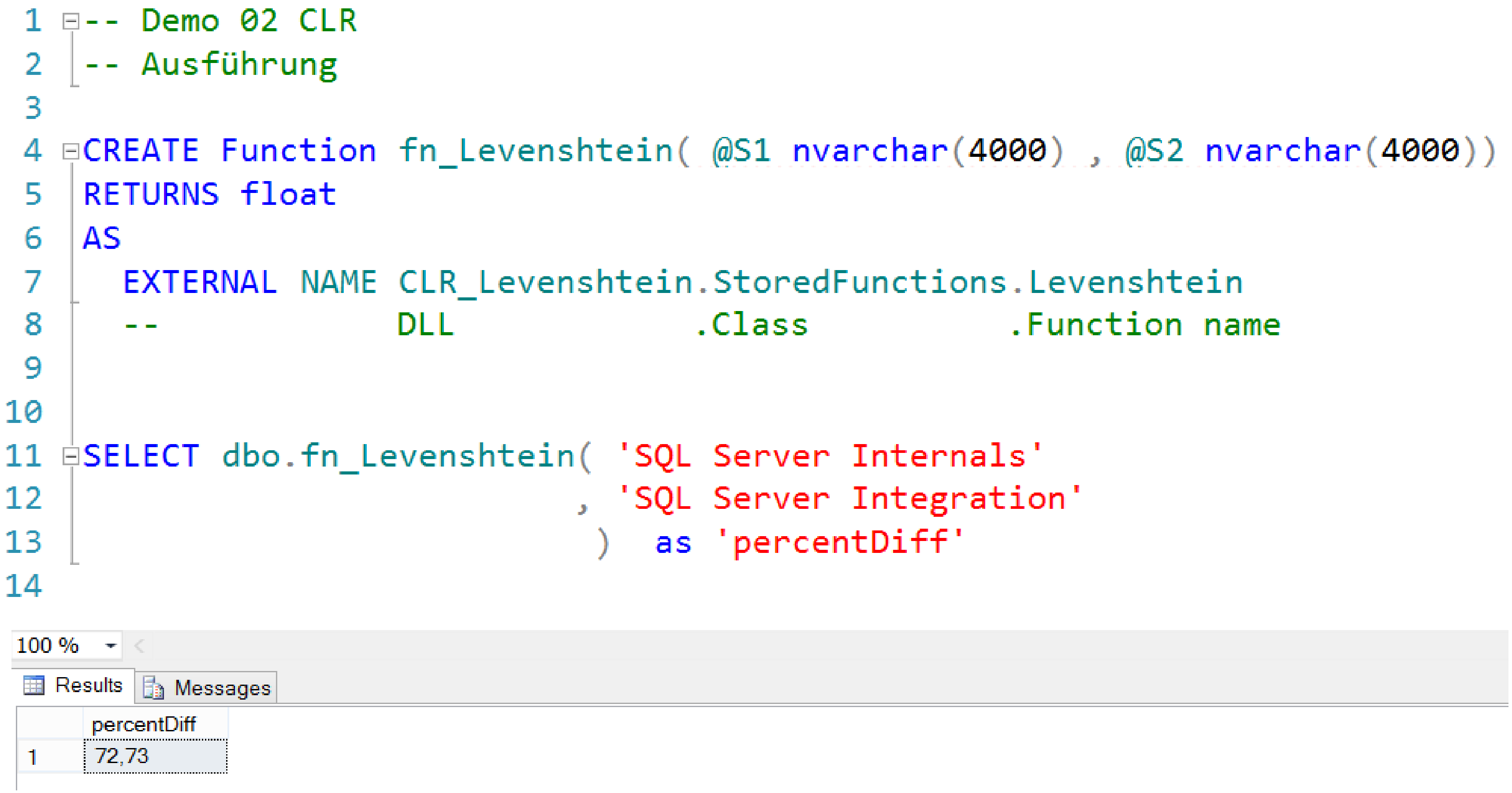
```
-- Demo 02 CLR
use test_CLR

-- aktivieren der CLR-Integration
EXEC sp_configure 'clr enabled', 1
go
reconfigure
go
---------

--Listing C# Assembly
using System;
using System.Data;
...

--> mit obigem Listing (xmlSaveProc.cs) muss mittels Compiler eine .dll erstellt werden.
 --  csc /t:library filename.cs
 --  evtl vorher path configurieren auf C:\WINDOWS\Microsoft.NET\Framework\v2.0.50727
go

-- Registrieren der Assembly
CREATE ASSEMBLY CLR_Levenshtein
FROM 'C:\ <folder> \CLR_Levenshtein.dll'
WITH Permission_Set = Safe

-- überprüfung
SELECT * FROM sys.assemblies
SELECT * FROM sys.assembly_files
```

Demo result



```
using System;
using System.Data;
using System.Data.SqlClient;
using System.Data.SqlTypes;
using Microsoft.SqlServer.Server;

public partial class StoredFunctions
{
    [Microsoft.SqlServer.Server.SqlFunction(IsDeterministic = true, IsPrecise = false)]
    public static SqlDouble Levenshtein(SqlString S1, SqlString S2)
    {
        if(S1.IsNull)
          S1 = new SqlString("");
        if(S2.IsNull)
          S2 = new SqlString("");
        String SC1 = S1.Value.ToUpper();
        String SC2 = S2.Value.ToUpper();
        int n = SC1.Length;
        int m = SC2.Length;

        int[,] d = new int[n + 1, m + 1];
        int cost = 0;

        if (n + m == 0) {
            return 100;
        } else if (n == 0) {
            return 0;
        } else if (m == 0) {
            return 0;
        }
        for (int i = 0; i <= n; i++)
            d[i, 0] = i;
        for (int j = 0; j <= m; j++)
            d[0, j] = j;
        for (int i = 1; i <= n; i++)
        {
            for (int j = 1; j <= m; j++)
            {
                if (SC1[i - 1] == SC2[j - 1])
                    cost = 0;
                else
                    cost = 1;
                d[i, j] = System.Math.Min(System.Math.Min(d[i - 1, j] + 1, d[i, j - 1] + 1), d[i - 1, j - 1] + cost);
            }
        }

        double percentage = System.Math.Round((1.0 - ((double)d[n, m]/(double)System.Math.Max(n,m))) * 100.0,2);
        return percentage;
    }
};
```

Demo result



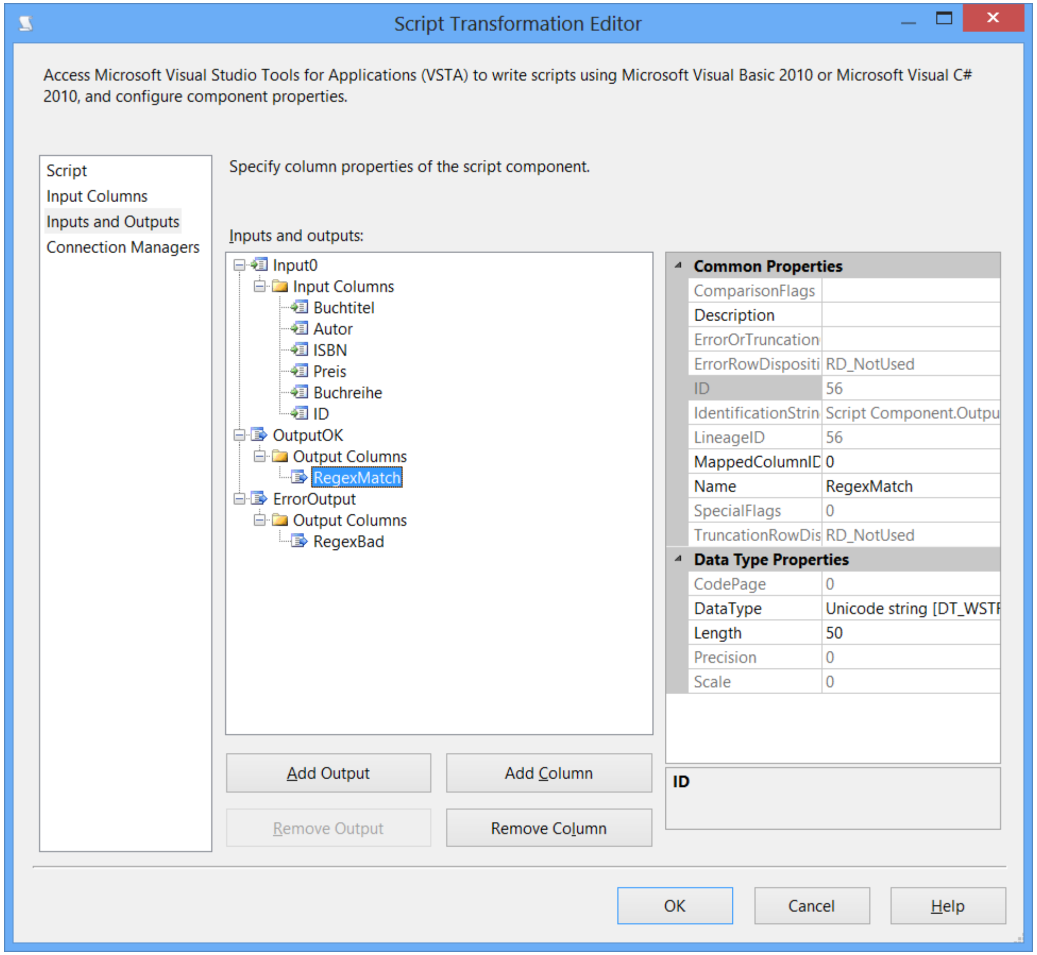
```sql
-- Demo 02 CLR
-- Ausführung

CREATE Function fn_Levenshtein( @S1 nvarchar(4000) , @S2 nvarchar(4000))
RETURNS float
AS
  EXTERNAL NAME CLR_Levenshtein.StoredFunctions.Levenshtein
  --            DLL            .Class          .Function name


SELECT dbo.fn_Levenshtein( 'SQL Server Internals'
                         , 'SQL Server Integration'
                          )  as 'percentDiff'

```

	percentDiff
1	72,73

SSIS custom components

- C# und VB .NET
- „frei programmierbar“
- synchron sehr performant
- optimale Verwendung im Dataflow
- konfigurierbar via „toolDB“

SSIS custom components



Script Transformation Editor

Access Microsoft Visual Studio Tools for Applications (VSTA) to write scripts using Microsoft Visual Basic 2010 or Microsoft Visual C# 2010, and configure component properties.

Specify column properties of the script component.

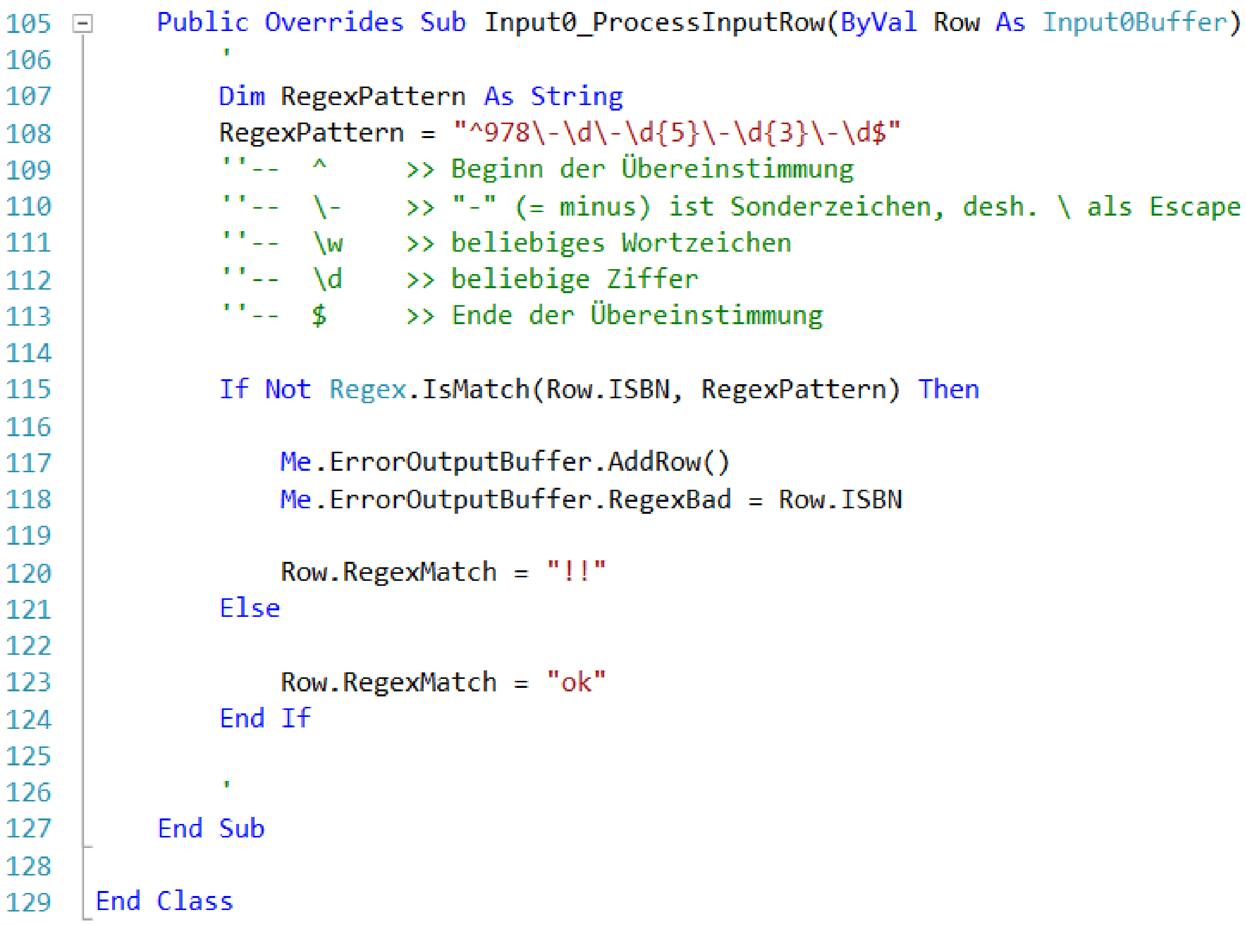
Inputs and outputs:

```
Input0
  Input Columns
    Buchtitel
    Autor
    ISBN
    Preis
    Buchreihe
    ID
OutputOK
  Output Columns
    RegexMatch
ErrorOutput
  Output Columns
    RegexBad
```

Common Properties	
ComparisonFlags	
Description	
ErrorOrTruncation	
ErrorRowDispositi	RD_NotUsed
ID	56
IdentificationStrin	Script Component.Outpu
LineageID	56
MappedColumnID	0
Name	RegexMatch
SpecialFlags	0
TruncationRowDis	RD_NotUsed
Data Type Properties	
CodePage	0
DataType	Unicode string [DT_WSTR
Length	50
Precision	0
Scale	0

ID

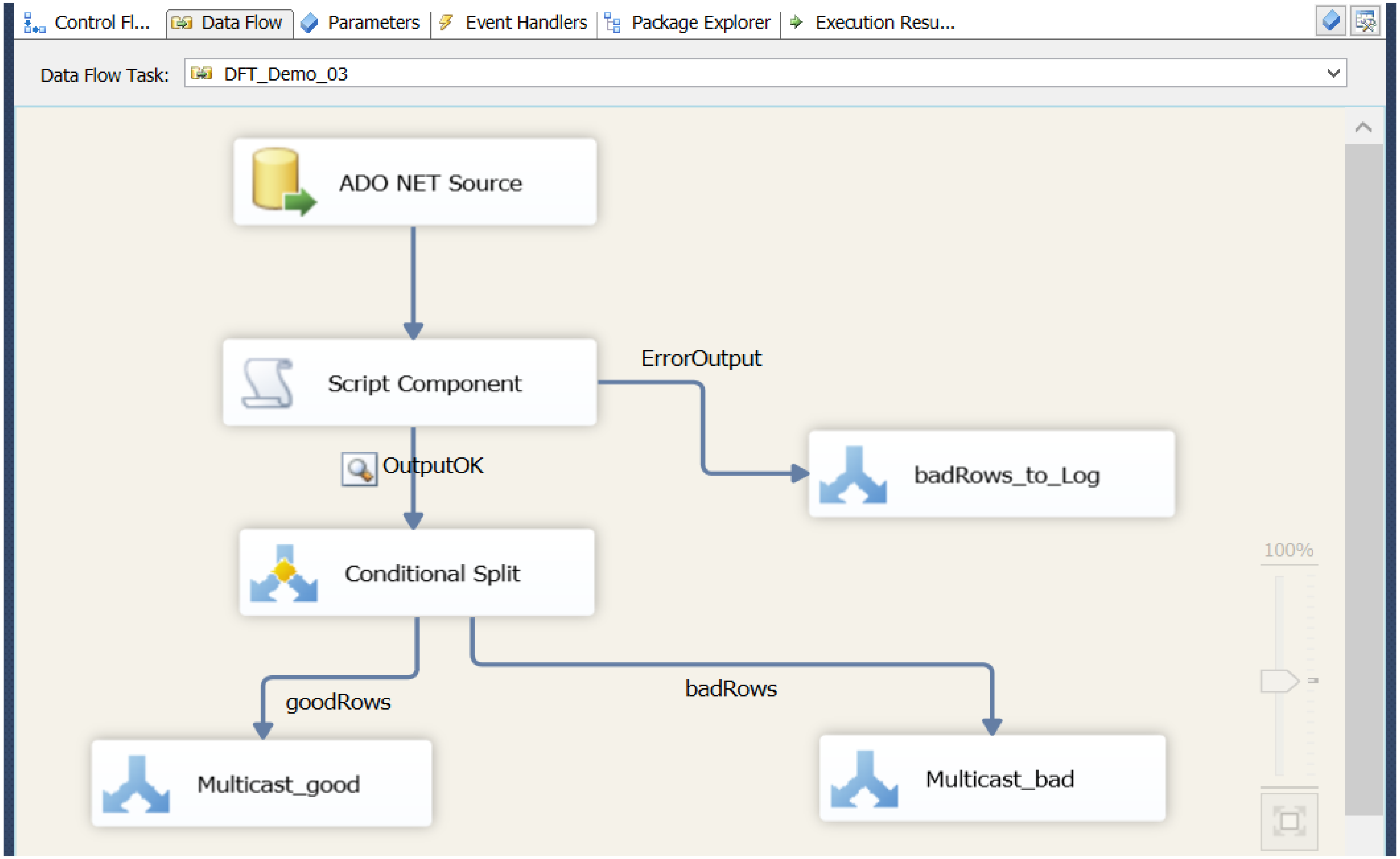
SSIS custom components



```
    Public Overrides Sub Input0_ProcessInputRow(ByVal Row As Input0Buffer)
        '
        Dim RegexPattern As String
        RegexPattern = "^978\-\d\-\d{5}\-\d{3}\-\d$"
        ''--  ^     >> Beginn der Übereinstimmung
        ''--  \-    >> "-" (= minus) ist Sonderzeichen, desh. \ als Escape
        ''--  \w    >> beliebiges Wortzeichen
        ''--  \d    >> beliebige Ziffer
        ''--  $     >> Ende der Übereinstimmung

        If Not Regex.IsMatch(Row.ISBN, RegexPattern) Then

            Me.ErrorOutputBuffer.AddRow()
            Me.ErrorOutputBuffer.RegexBad = Row.ISBN

            Row.RegexMatch = "!!"
        Else

            Row.RegexMatch = "ok"
        End If

        '
    End Sub

End Class
```

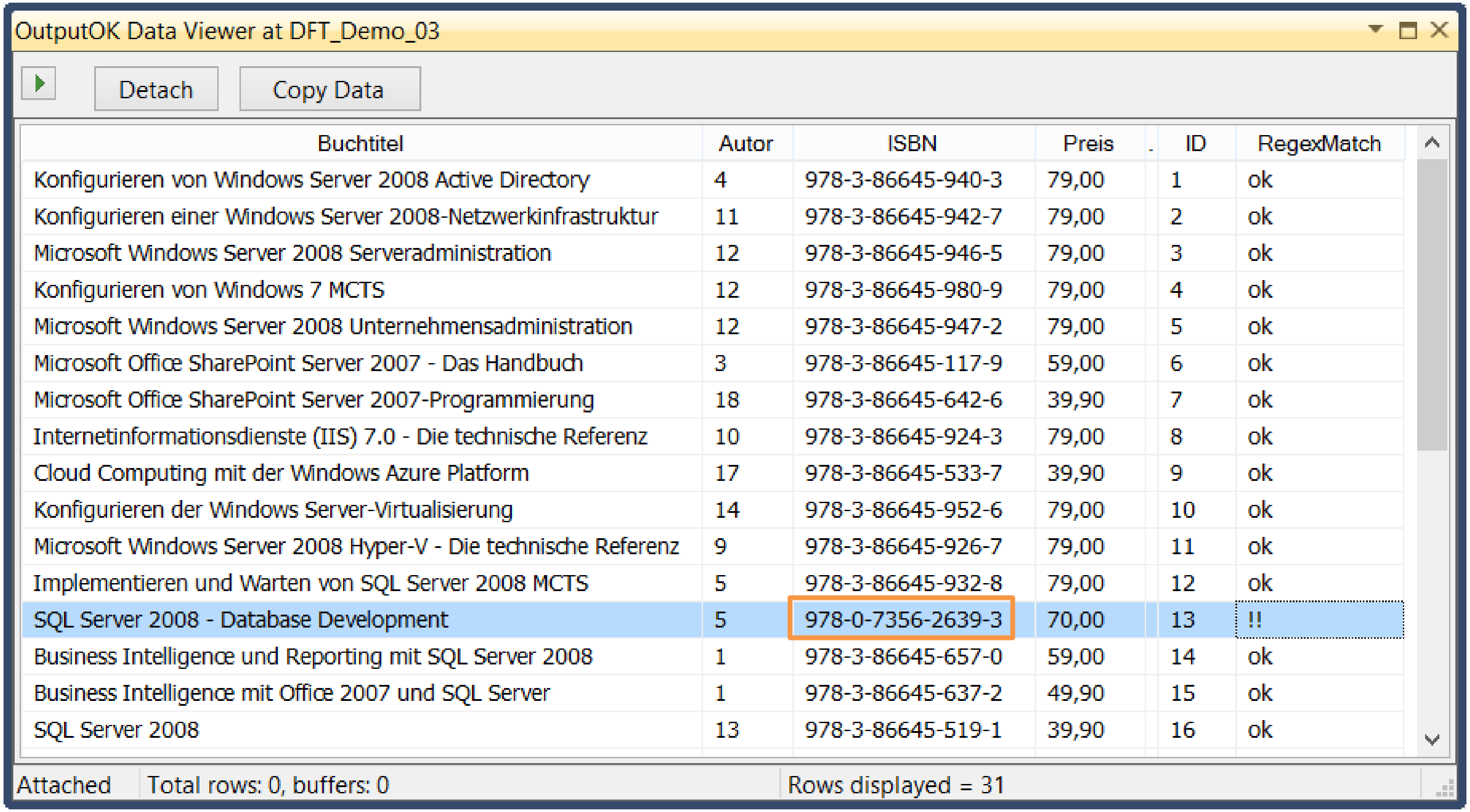
DEMO

Demo result



Data Flow Task: DFT_Demo_03

Demo result



Buchtitel	Autor	ISBN	Preis	.	ID	RegexMatch
Konfigurieren von Windows Server 2008 Active Directory	4	978-3-86645-940-3	79,00		1	ok
Konfigurieren einer Windows Server 2008-Netzwerkinfrastruktur	11	978-3-86645-942-7	79,00		2	ok
Microsoft Windows Server 2008 Serveradministration	12	978-3-86645-946-5	79,00		3	ok
Konfigurieren von Windows 7 MCTS	12	978-3-86645-980-9	79,00		4	ok
Microsoft Windows Server 2008 Unternehmensadministration	12	978-3-86645-947-2	79,00		5	ok
Microsoft Office SharePoint Server 2007 - Das Handbuch	3	978-3-86645-117-9	59,00		6	ok
Microsoft Office SharePoint Server 2007-Programmierung	18	978-3-86645-642-6	39,90		7	ok
Internetinformationsdienste (IIS) 7.0 - Die technische Referenz	10	978-3-86645-924-3	79,00		8	ok
Cloud Computing mit der Windows Azure Platform	17	978-3-86645-533-7	39,90		9	ok
Konfigurieren der Windows Server-Virtualisierung	14	978-3-86645-952-6	79,00		10	ok
Microsoft Windows Server 2008 Hyper-V - Die technische Referenz	9	978-3-86645-926-7	79,00		11	ok
Implementieren und Warten von SQL Server 2008 MCTS	5	978-3-86645-932-8	79,00		12	ok
SQL Server 2008 - Database Development	5	978-0-7356-2639-3	70,00		13	!!
Business Intelligence und Reporting mit SQL Server 2008	1	978-3-86645-657-0	59,00		14	ok
Business Intelligence mit Office 2007 und SQL Server	1	978-3-86645-637-2	49,90		15	ok
SQL Server 2008	13	978-3-86645-519-1	39,90		16	ok

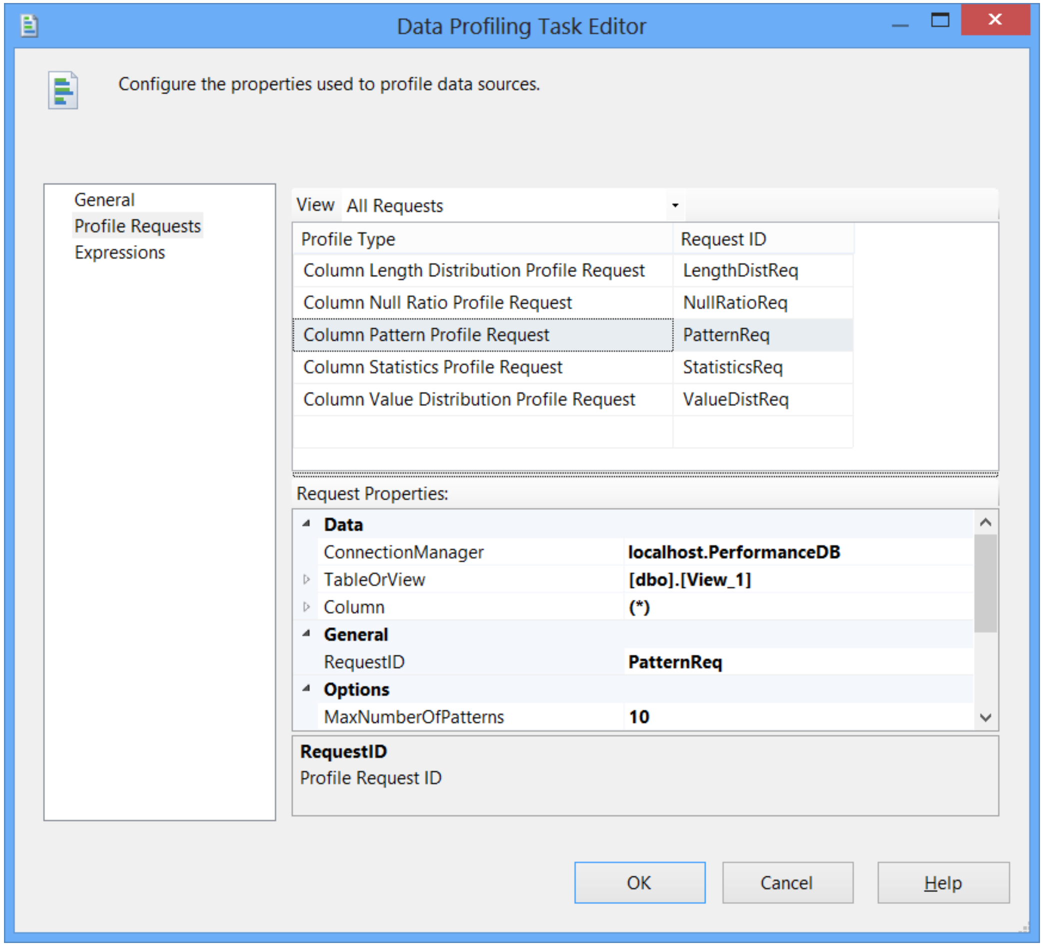
Attached | Total rows: 0, buffers: 0 | Rows displayed = 31

SSIS

Data Profiling Task

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

SSIS Data Profiling Task



Data Profiling Task Editor

Configure the properties used to profile data sources.

View: All Requests

Profile Type	Request ID
Column Length Distribution Profile Request	LengthDistReq
Column Null Ratio Profile Request	NullRatioReq
Column Pattern Profile Request	PatternReq
Column Statistics Profile Request	StatisticsReq
Column Value Distribution Profile Request	ValueDistReq

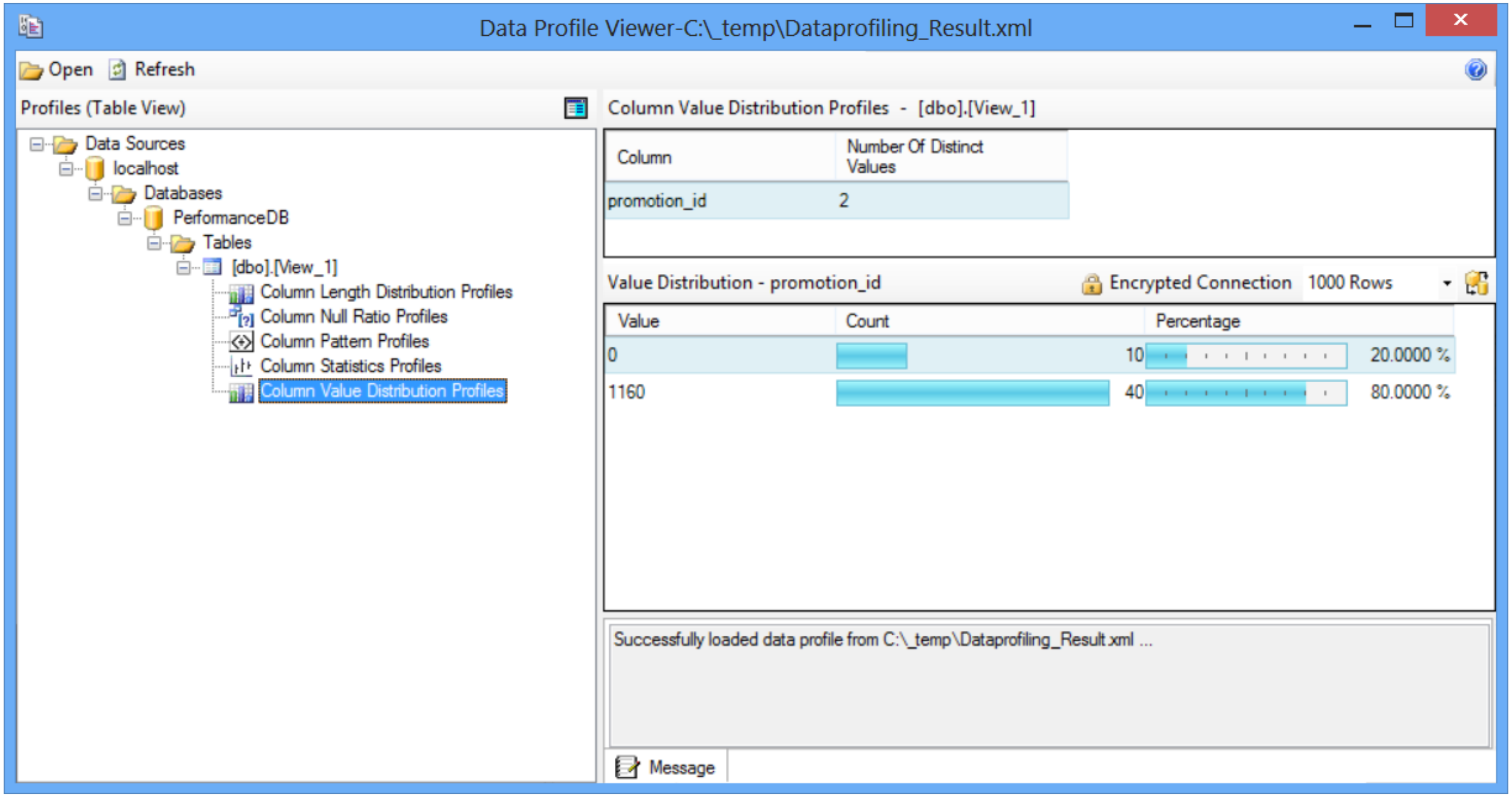
Request Properties:

Data	
ConnectionManager	**localhost.PerformanceDB**
TableOrView	**[dbo].[View_1]**
Column	**(*)**
General	
RequestID	**PatternReq**
Options	
MaxNumberOfPatterns	**10**

RequestID
Profile Request ID

DEMO

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
Encrypted Connection
1000 Rows
Message

Column Value Distribution Profiles - [dbo].[View_1]

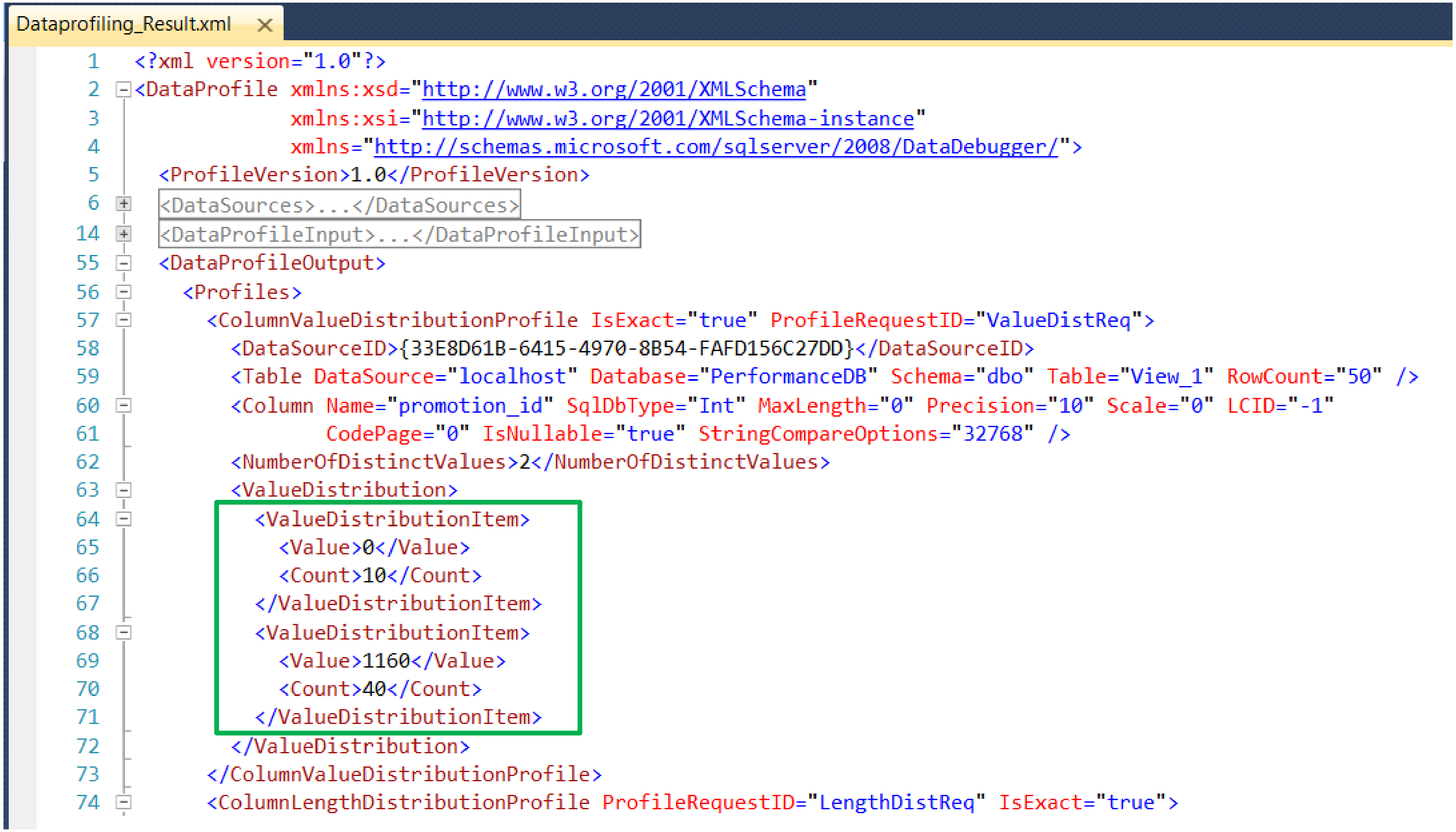
Column	Number Of Distinct Values
promotion_id	2

Value Distribution - promotion_id

Value	Count	Percentage
0	10	20.0000 %
1160	40	80.0000 %

Successfully loaded data profile from C:_temp\Dataprofiling_Result.xml ...

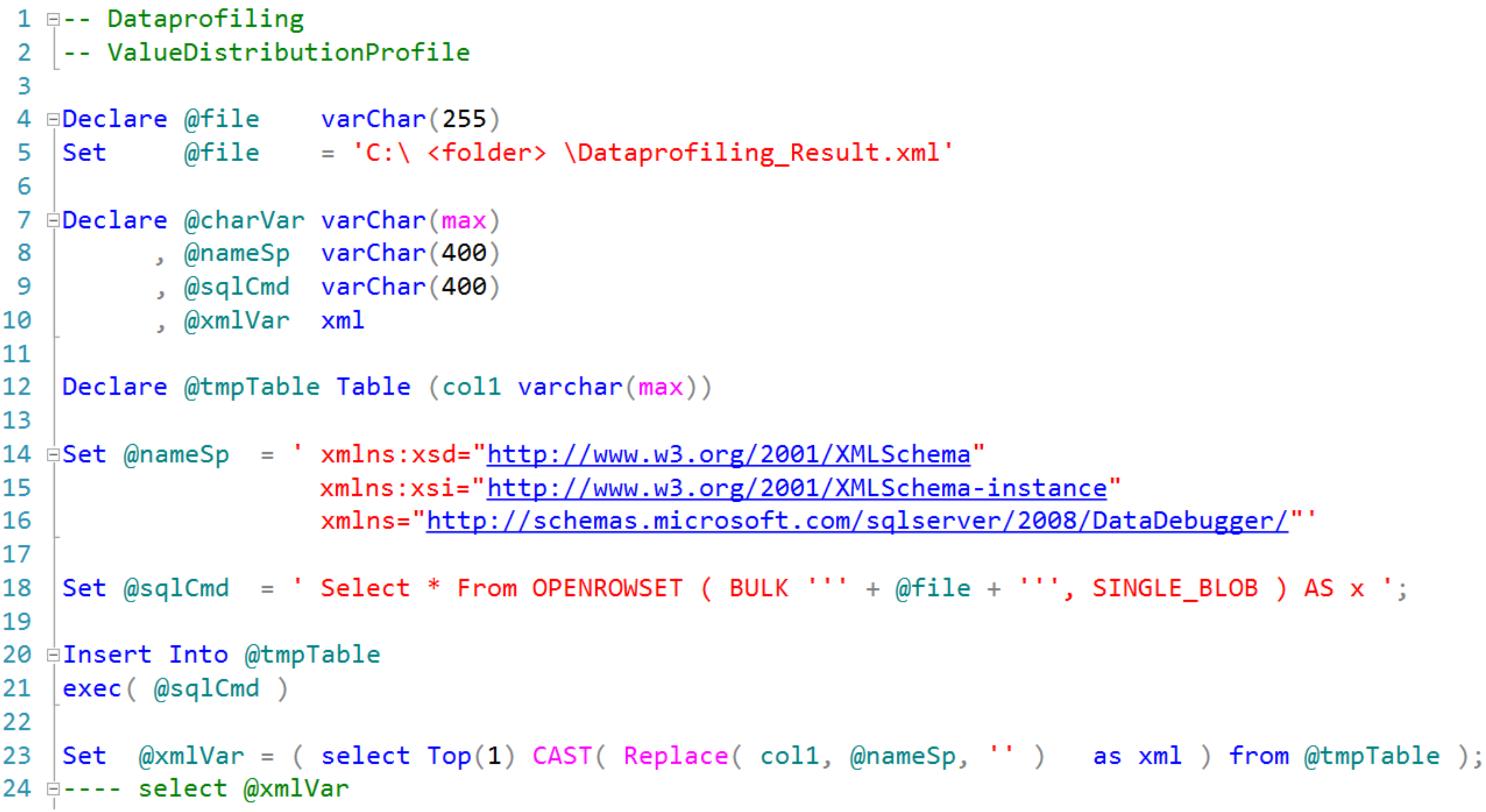
Demo result



Dataprofiling_Result.xml

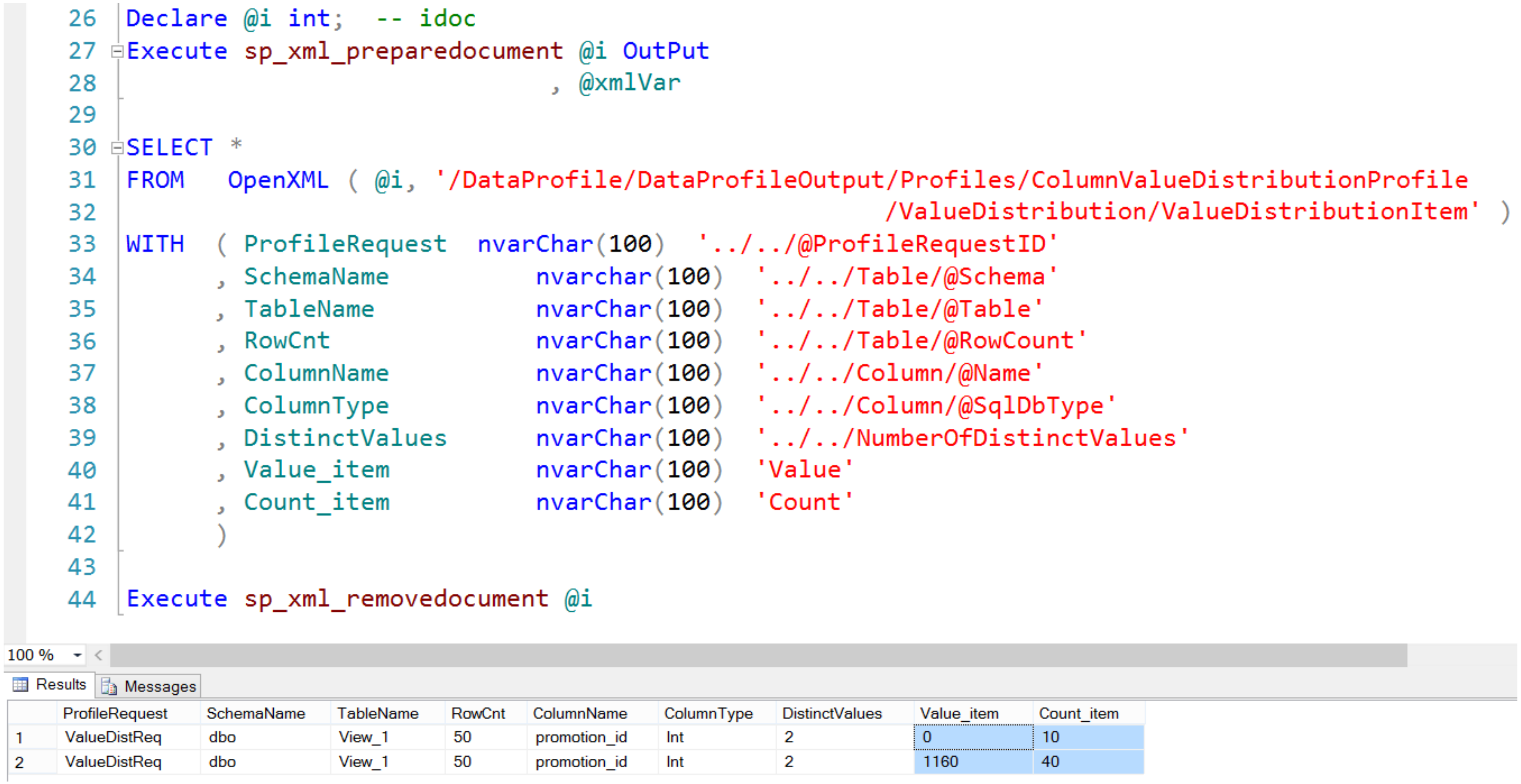
```xml
<?xml version="1.0"?>
<DataProfile xmlns:xsd="http://www.w3.org/2001/XMLSchema"
             xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
             xmlns="http://schemas.microsoft.com/sqlserver/2008/DataDebugger/">
  <ProfileVersion>1.0</ProfileVersion>
  <DataSources>...</DataSources>
  <DataProfileInput>...</DataProfileInput>
  <DataProfileOutput>
    <Profiles>
      <ColumnValueDistributionProfile IsExact="true" ProfileRequestID="ValueDistReq">
        <DataSourceID>{33E8D61B-6415-4970-8B54-FAFD156C27DD}</DataSourceID>
        <Table DataSource="localhost" Database="PerformanceDB" Schema="dbo" Table="View_1" RowCount="50" />
        <Column Name="promotion_id" SqlDbType="Int" MaxLength="0" Precision="10" Scale="0" LCID="-1"
                CodePage="0" IsNullable="true" StringCompareOptions="32768" />
        <NumberOfDistinctValues>2</NumberOfDistinctValues>
        <ValueDistribution>
          <ValueDistributionItem>
            <Value>0</Value>
            <Count>10</Count>
          </ValueDistributionItem>
          <ValueDistributionItem>
            <Value>1160</Value>
            <Count>40</Count>
          </ValueDistributionItem>
        </ValueDistribution>
      </ColumnValueDistributionProfile>
      <ColumnLengthDistributionProfile ProfileRequestID="LengthDistReq" IsExact="true">
```

Demo result



```
-- Dataprofiling
-- ValueDistributionProfile

Declare @file    varChar(255)
Set     @file    = 'C:\ <folder> \Dataprofiling_Result.xml'

Declare @charVar varChar(max)
      , @nameSp  varChar(400)
      , @sqlCmd  varChar(400)
      , @xmlVar  xml

Declare @tmpTable Table (col1 varchar(max))

Set @nameSp  = ' xmlns:xsd="http://www.w3.org/2001/XMLSchema"
                 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                 xmlns="http://schemas.microsoft.com/sqlserver/2008/DataDebugger/"'

Set @sqlCmd  = ' Select * From OPENROWSET ( BULK ''' + @file + ''', SINGLE_BLOB ) AS x ';

Insert Into @tmpTable
exec( @sqlCmd )

Set  @xmlVar = ( select Top(1) CAST( Replace( col1, @nameSp, '' )   as xml ) from @tmpTable );
---- select @xmlVar
```

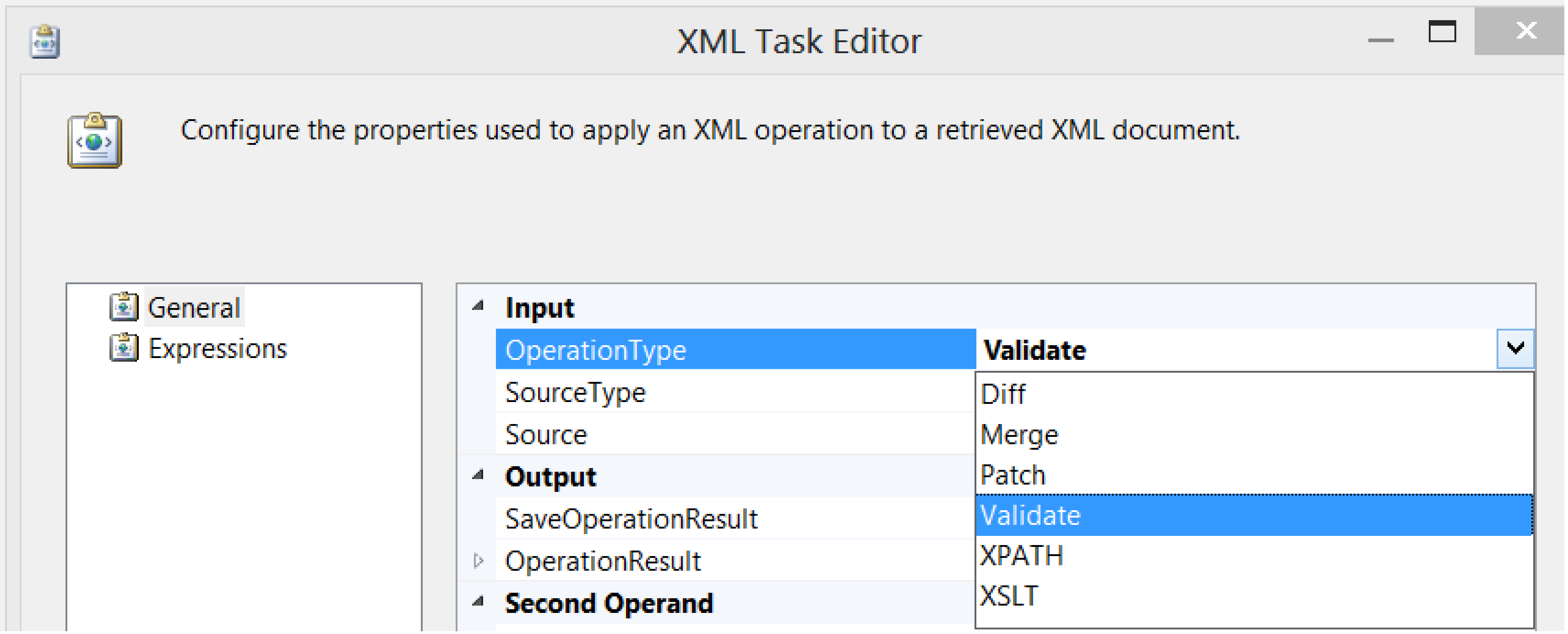
Demo result



```
Declare @i int;   -- idoc
Execute sp_xml_preparedocument @i OutPut
                             , @xmlVar

SELECT *
FROM   OpenXML ( @i, '/DataProfile/DataProfileOutput/Profiles/ColumnValueDistributionProfile
                                                    /ValueDistribution/ValueDistributionItem' )
WITH   ( ProfileRequest  nvarChar(100)  '../../@ProfileRequestID'
       , SchemaName          nvarchar(100)  '../../Table/@Schema'
       , TableName           nvarChar(100)  '../../Table/@Table'
       , RowCnt              nvarChar(100)  '../../Table/@RowCount'
       , ColumnName          nvarChar(100)  '../../Column/@Name'
       , ColumnType          nvarChar(100)  '../../Column/@SqlDbType'
       , DistinctValues      nvarChar(100)  '../../NumberOfDistinctValues'
       , Value_item          nvarChar(100)  'Value'
       , Count_item          nvarChar(100)  'Count'
       )

Execute sp_xml_removedocument @i
```

	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

XML Task


XML Task Editor

Configure the properties used to apply an XML operation to a retrieved XML document.

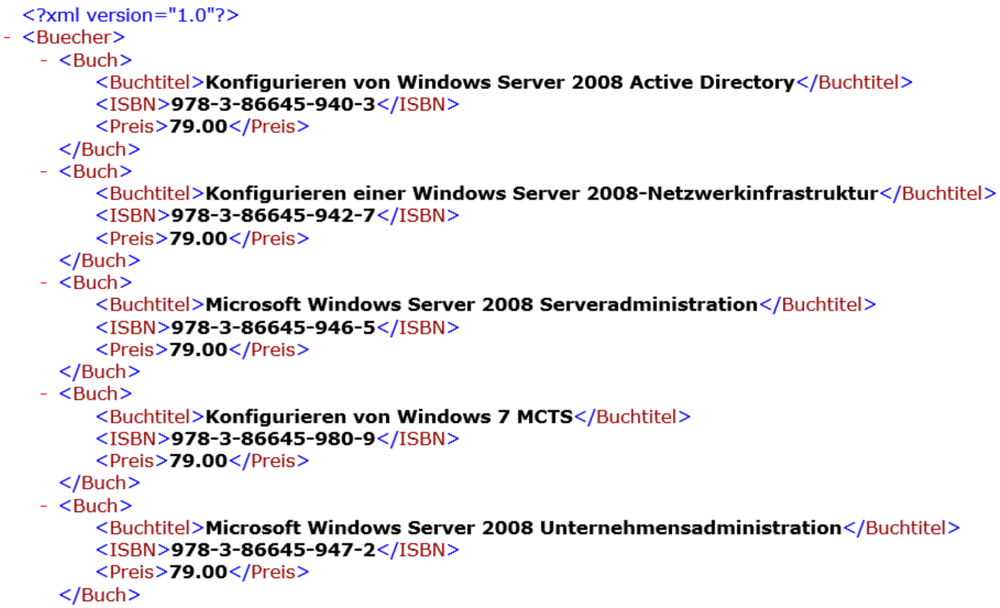
Input

OperationType	**Validate**
SourceType	
Source	

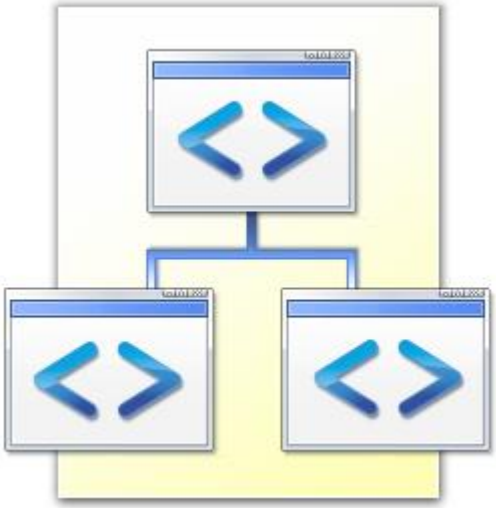
Output

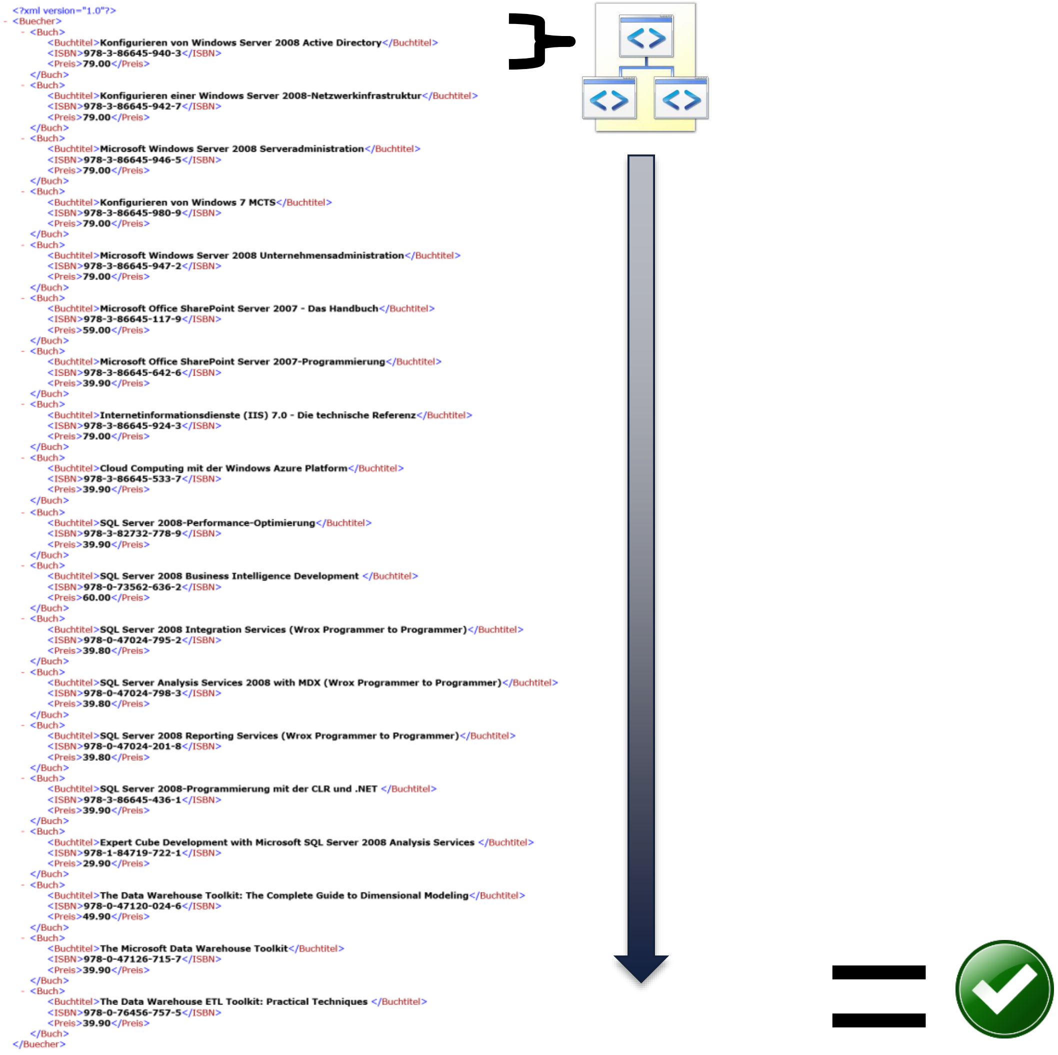
SaveOperationResult	
OperationResult	

Second Operand



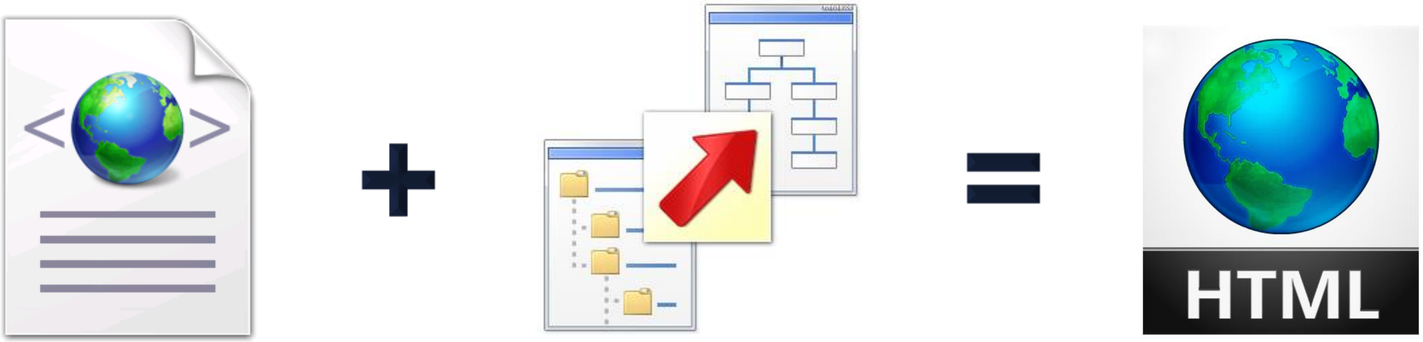
```
  <?xml version="1.0"?>
- <Buecher>
  - <Buch>
        <Buchtitel>Konfigurieren von Windows Server 2008 Active Directory</Buchtitel>
        <ISBN>978-3-86645-940-3</ISBN>
        <Preis>79.00</Preis>
    </Buch>
  - <Buch>
        <Buchtitel>Konfigurieren einer Windows Server 2008-Netzwerkinfrastruktur</Buchtitel>
        <ISBN>978-3-86645-942-7</ISBN>
        <Preis>79.00</Preis>
    </Buch>
  - <Buch>
        <Buchtitel>Microsoft Windows Server 2008 Serveradministration</Buchtitel>
        <ISBN>978-3-86645-946-5</ISBN>
        <Preis>79.00</Preis>
    </Buch>
  - <Buch>
        <Buchtitel>Konfigurieren von Windows 7 MCTS</Buchtitel>
        <ISBN>978-3-86645-980-9</ISBN>
        <Preis>79.00</Preis>
    </Buch>
  - <Buch>
        <Buchtitel>Microsoft Windows Server 2008 Unternehmensadministration</Buchtitel>
        <ISBN>978-3-86645-947-2</ISBN>
        <Preis>79.00</Preis>
    </Buch>
```





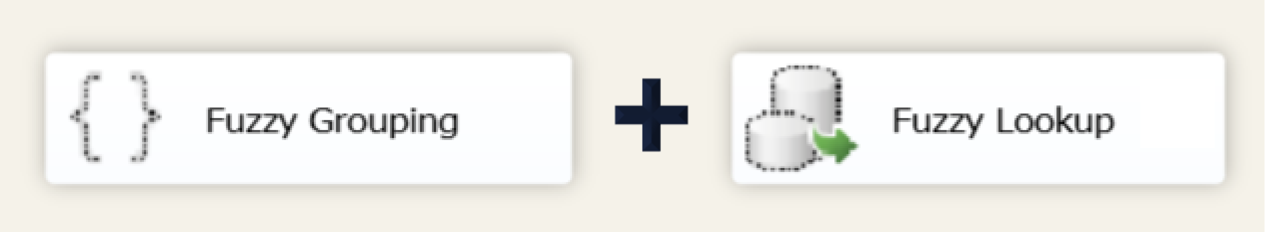
```
  <?xml version="1.0"?>
- <Buecher>
  - <Buch>
      <Buchtitel>Konfigurieren von Windows Server 2008 Active Directory</Buchtitel>
      <ISBN>978-3-86645-940-3</ISBN>
      <Preis>79.00</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>Konfigurieren einer Windows Server 2008-Netzwerkinfrastruktur</Buchtitel>
      <ISBN>978-3-86645-942-7</ISBN>
      <Preis>79.00</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>Microsoft Windows Server 2008 Serveradministration</Buchtitel>
      <ISBN>978-3-86645-946-5</ISBN>
      <Preis>79.00</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>Konfigurieren von Windows 7 MCTS</Buchtitel>
      <ISBN>978-3-86645-980-9</ISBN>
      <Preis>79.00</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>Microsoft Windows Server 2008 Unternehmensadministration</Buchtitel>
      <ISBN>978-3-86645-947-2</ISBN>
      <Preis>79.00</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>Microsoft Office SharePoint Server 2007 - Das Handbuch</Buchtitel>
      <ISBN>978-3-86645-117-9</ISBN>
      <Preis>59.00</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>Microsoft Office SharePoint Server 2007-Programmierung</Buchtitel>
      <ISBN>978-3-86645-642-6</ISBN>
      <Preis>39.90</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>Internetinformationsdienste (IIS) 7.0 - Die technische Referenz</Buchtitel>
      <ISBN>978-3-86645-924-3</ISBN>
      <Preis>79.00</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>Cloud Computing mit der Windows Azure Platform</Buchtitel>
      <ISBN>978-3-86645-533-7</ISBN>
      <Preis>39.90</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>SQL Server 2008-Performance-Optimierung</Buchtitel>
      <ISBN>978-3-82732-778-9</ISBN>
      <Preis>39.90</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>SQL Server 2008 Business Intelligence Development </Buchtitel>
      <ISBN>978-0-73562-636-2</ISBN>
      <Preis>60.00</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>SQL Server 2008 Integration Services (Wrox Programmer to Programmer)</Buchtitel>
      <ISBN>978-0-47024-795-2</ISBN>
      <Preis>39.80</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>SQL Server Analysis Services 2008 with MDX (Wrox Programmer to Programmer)</Buchtitel>
      <ISBN>978-0-47024-798-3</ISBN>
      <Preis>39.80</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>SQL Server 2008 Reporting Services (Wrox Programmer to Programmer)</Buchtitel>
      <ISBN>978-0-47024-201-8</ISBN>
      <Preis>39.80</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>SQL Server 2008-Programmierung mit der CLR und .NET </Buchtitel>
      <ISBN>978-3-86645-436-1</ISBN>
      <Preis>39.90</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>Expert Cube Development with Microsoft SQL Server 2008 Analysis Services </Buchtitel>
      <ISBN>978-1-84719-722-1</ISBN>
      <Preis>29.90</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling</Buchtitel>
      <ISBN>978-0-47120-024-6</ISBN>
      <Preis>49.90</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>The Microsoft Data Warehouse Toolkit</Buchtitel>
      <ISBN>978-0-47126-715-7</ISBN>
      <Preis>39.90</Preis>
    </Buch>
  - <Buch>
      <Buchtitel>The Data Warehouse ETL Toolkit: Practical Techniques </Buchtitel>
      <ISBN>978-0-76456-757-5</ISBN>
      <Preis>39.90</Preis>
    </Buch>
  </Buecher>
```

XML Task



DEMO

Fuzzy Grouping + Lookup

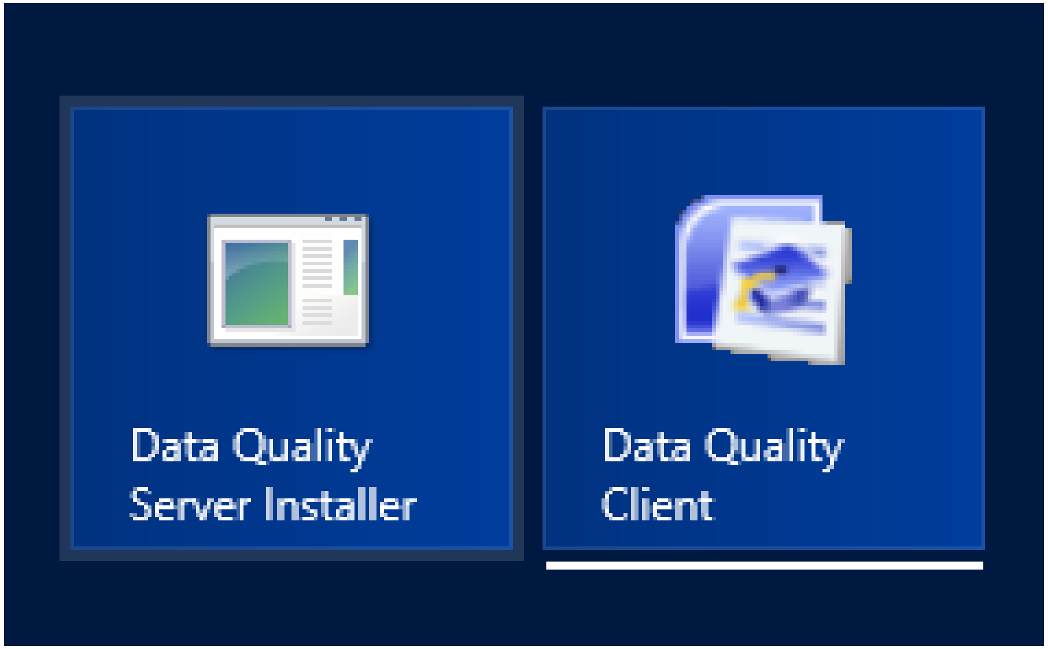


bisherige Lösungen

- Nachteil „handmade“
- ggf. Portierungsaufwand
- umfangreichere Doku
- Benutzer-Einweisung in workflow

>> DQS (neu ab 2012)

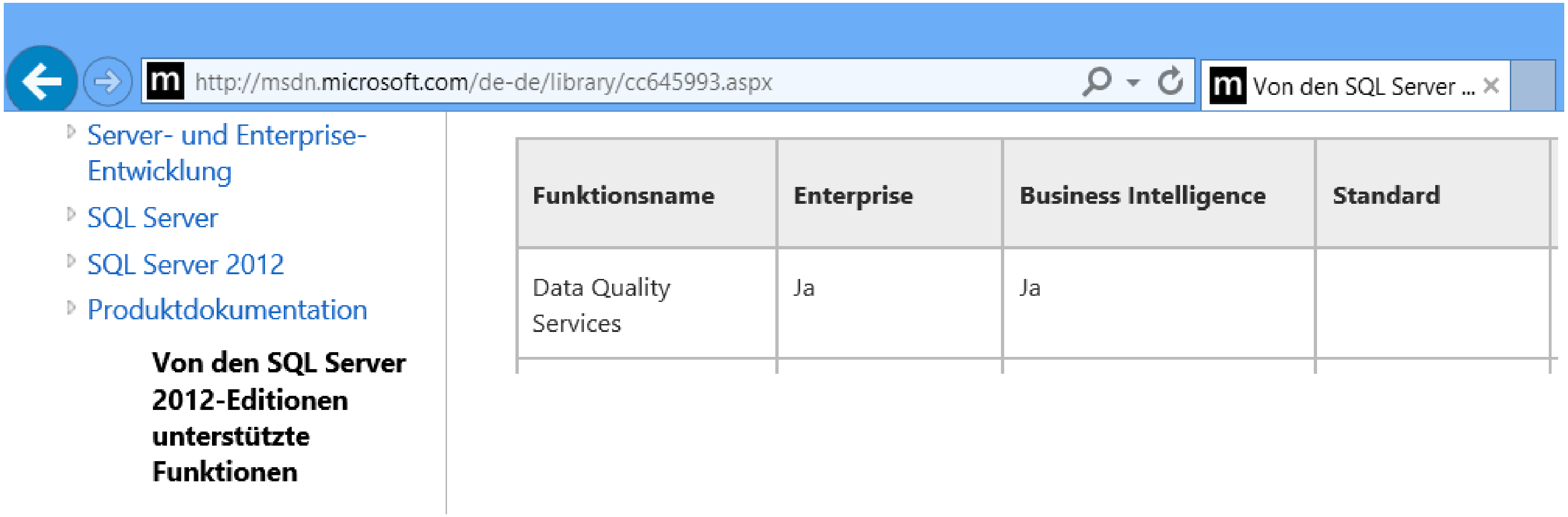
DQS Data Quality Service



Data Quality Server Installer

Data Quality Client

SQL Server 2012 Editionen

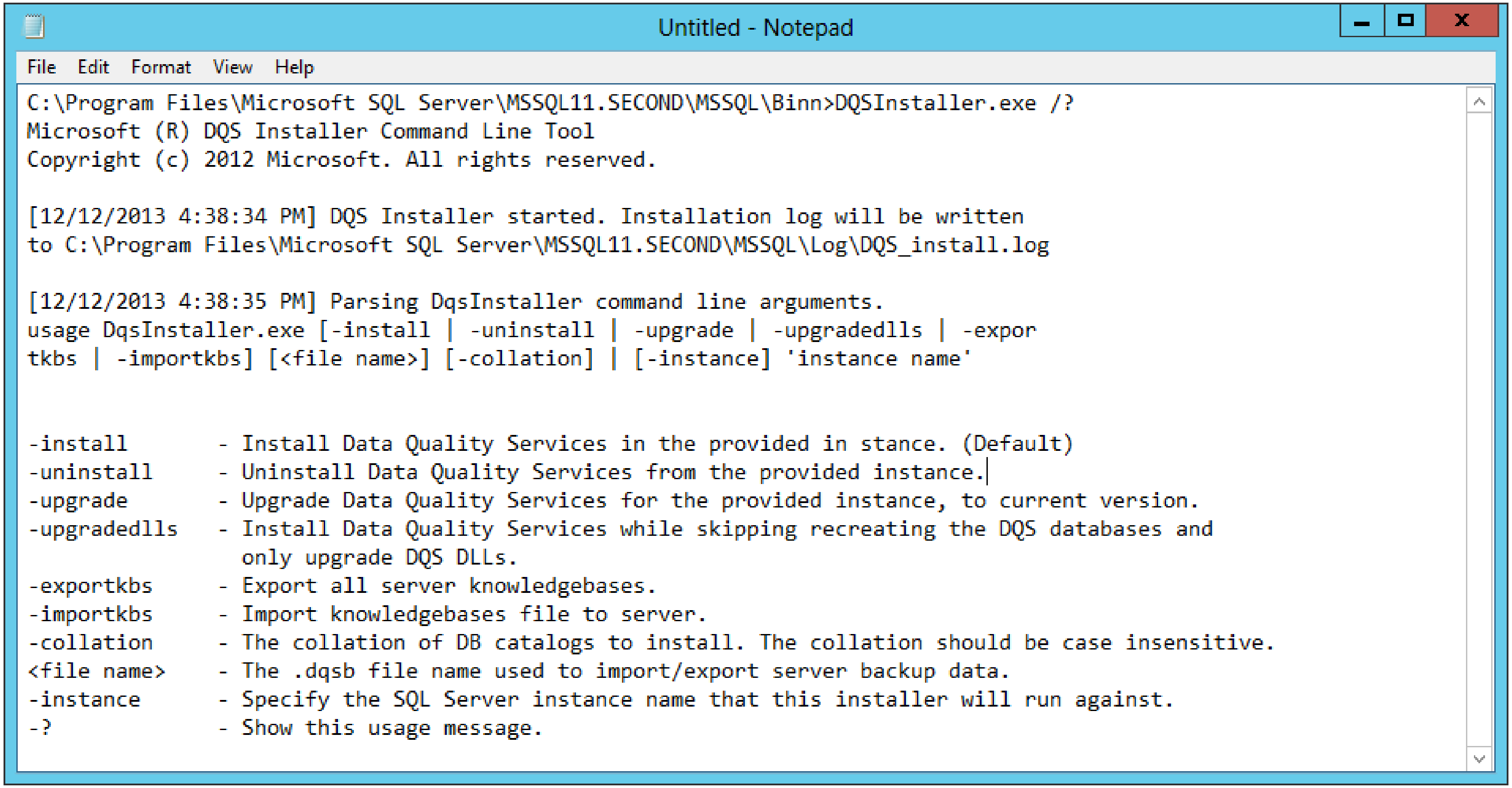


- Server- und Enterprise-Entwicklung
- SQL Server
- SQL Server 2012
- Produktdokumentation

Von den SQL Server 2012-Editionen unterstützte Funktionen

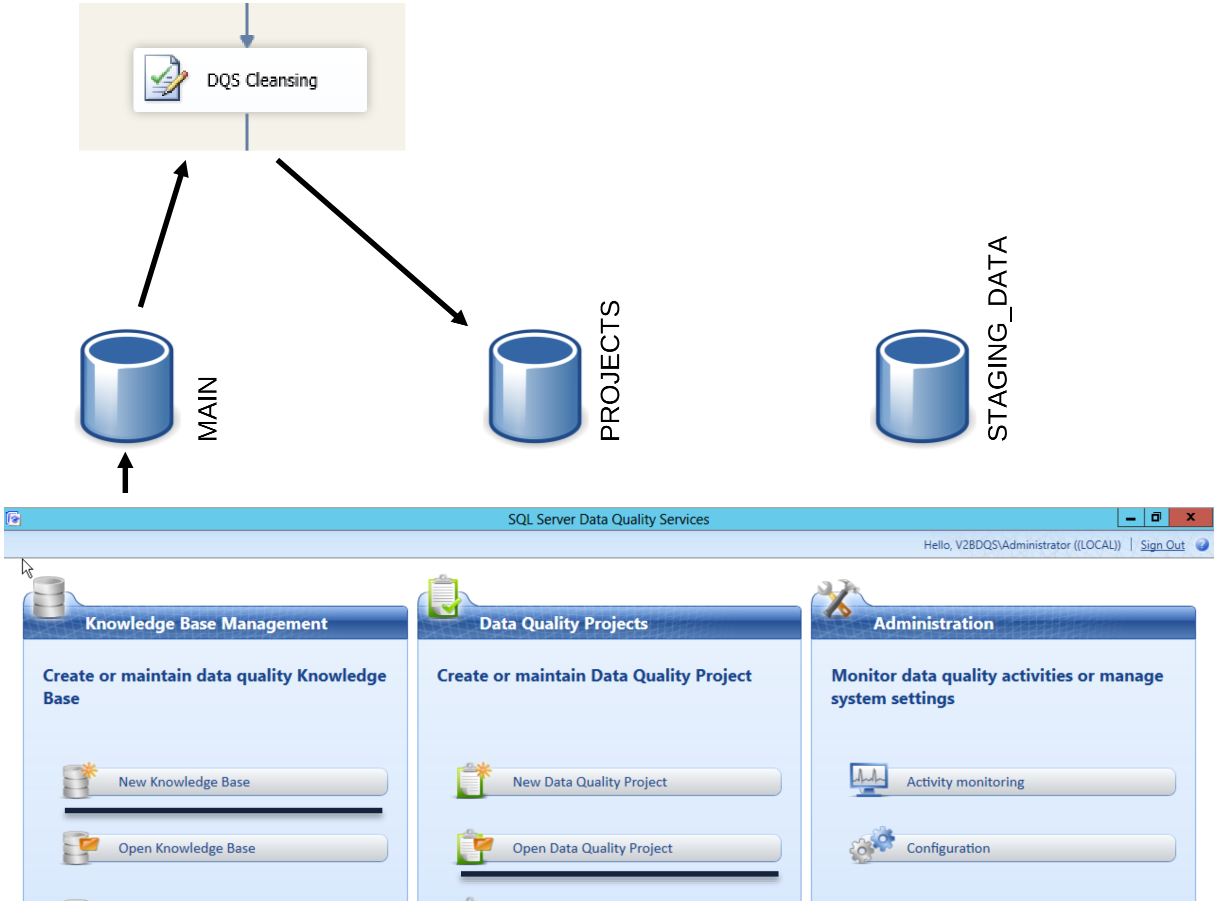
Funktionsname	Enterprise	Business Intelligence	Standard
Data Quality Services	Ja	Ja	

DQS Installer



```
C:\Program Files\Microsoft SQL Server\MSSQL11.SECOND\MSSQL\Binn>DQSInstaller.exe /?
Microsoft (R) DQS Installer Command Line Tool
Copyright (c) 2012 Microsoft. All rights reserved.

[12/12/2013 4:38:34 PM] DQS Installer started. Installation log will be written
to C:\Program Files\Microsoft SQL Server\MSSQL11.SECOND\MSSQL\Log\DQS_install.log

[12/12/2013 4:38:35 PM] Parsing DqsInstaller command line arguments.
usage DqsInstaller.exe [-install | -uninstall | -upgrade | -upgradedlls | -expor
tkbs | -importkbs] [<file name>] [-collation] | [-instance] 'instance name'


-install       - Install Data Quality Services in the provided in stance. (Default)
-uninstall     - Uninstall Data Quality Services from the provided instance.
-upgrade       - Upgrade Data Quality Services for the provided instance, to current version.
-upgradedlls   - Install Data Quality Services while skipping recreating the DQS databases and
                 only upgrade DQS DLLs.
-exportkbs     - Export all server knowledgebases.
-importkbs     - Import knowledgebases file to server.
-collation     - The collation of DB catalogs to install. The collation should be case insensitive.
<file name>    - The .dqsb file name used to import/export server backup data.
-instance      - Specify the SQL Server instance name that this installer will run against.
-?             - Show this usage message.
```

DQS Übersicht

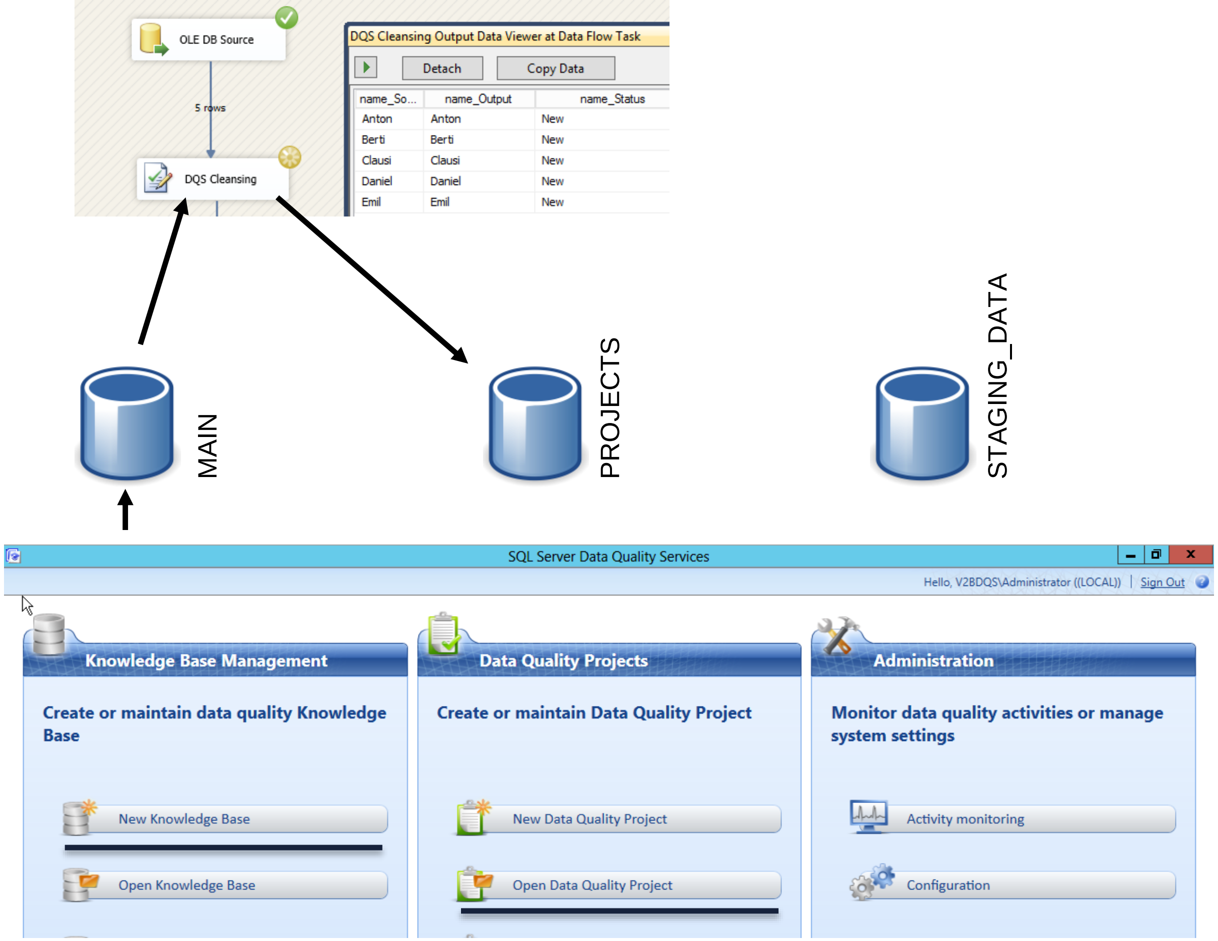


Create or maintain data quality Knowledge Base

Create or maintain Data Quality Project

Monitor data quality activities or manage system settings

DQS Übersicht



DQS Cleansing Output Data Viewer at Data Flow Task

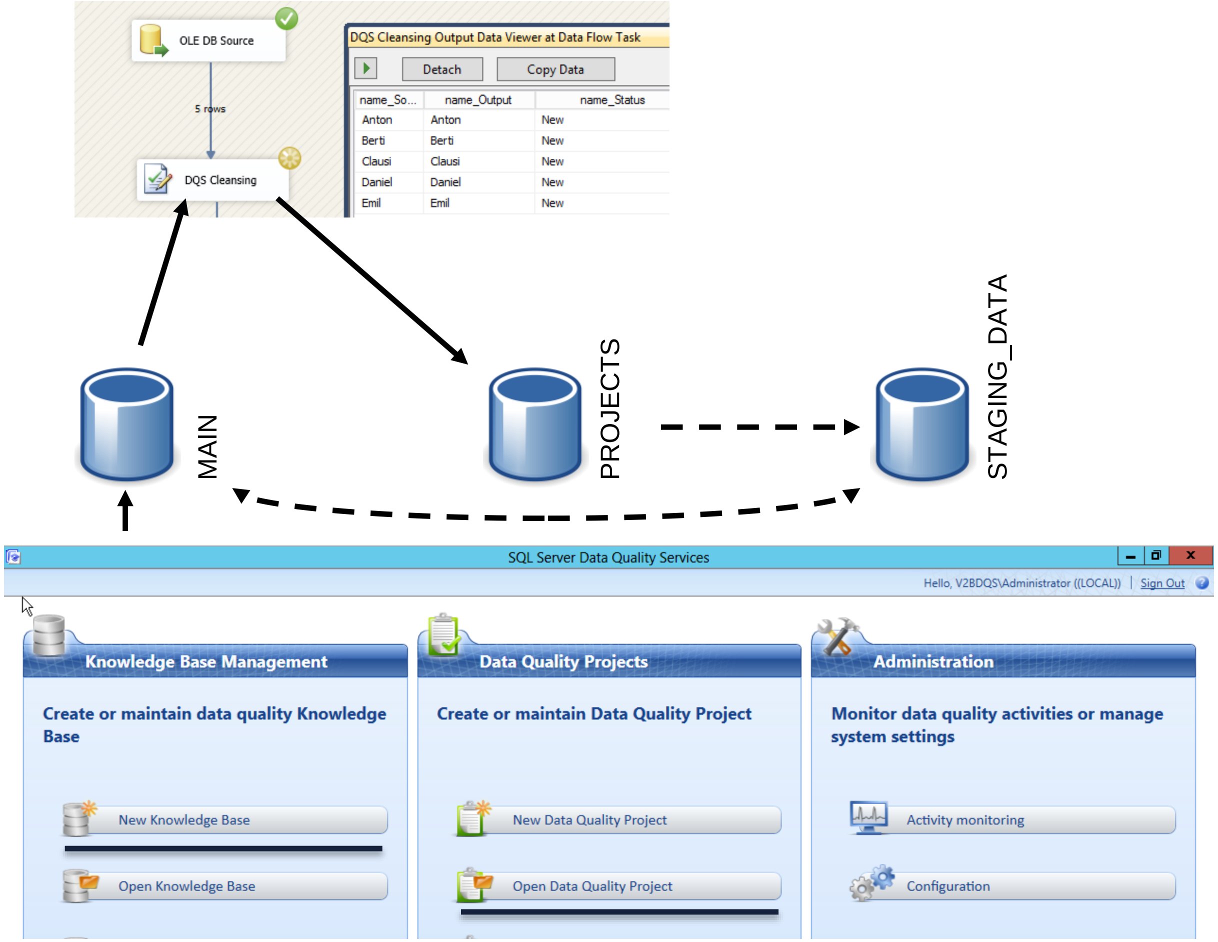
name_So...	name_Output	name_Status
Anton	Anton	New
Berti	Berti	New
Clausi	Clausi	New
Daniel	Daniel	New
Emil	Emil	New

Create or maintain data quality Knowledge Base

Create or maintain Data Quality Project

Monitor data quality activities or manage system settings

DQS Übersicht



DQS Cleansing Output Data Viewer at Data Flow Task

name_So...	name_Output	name_Status
Anton	Anton	New
Berti	Berti	New
Clausi	Clausi	New
Daniel	Daniel	New
Emil	Emil	New

SQL Server Data Quality Services

Hello, V2BDQS\Administrator ((LOCAL)) | Sign Out

Knowledge Base Management

Create or maintain data quality Knowledge Base

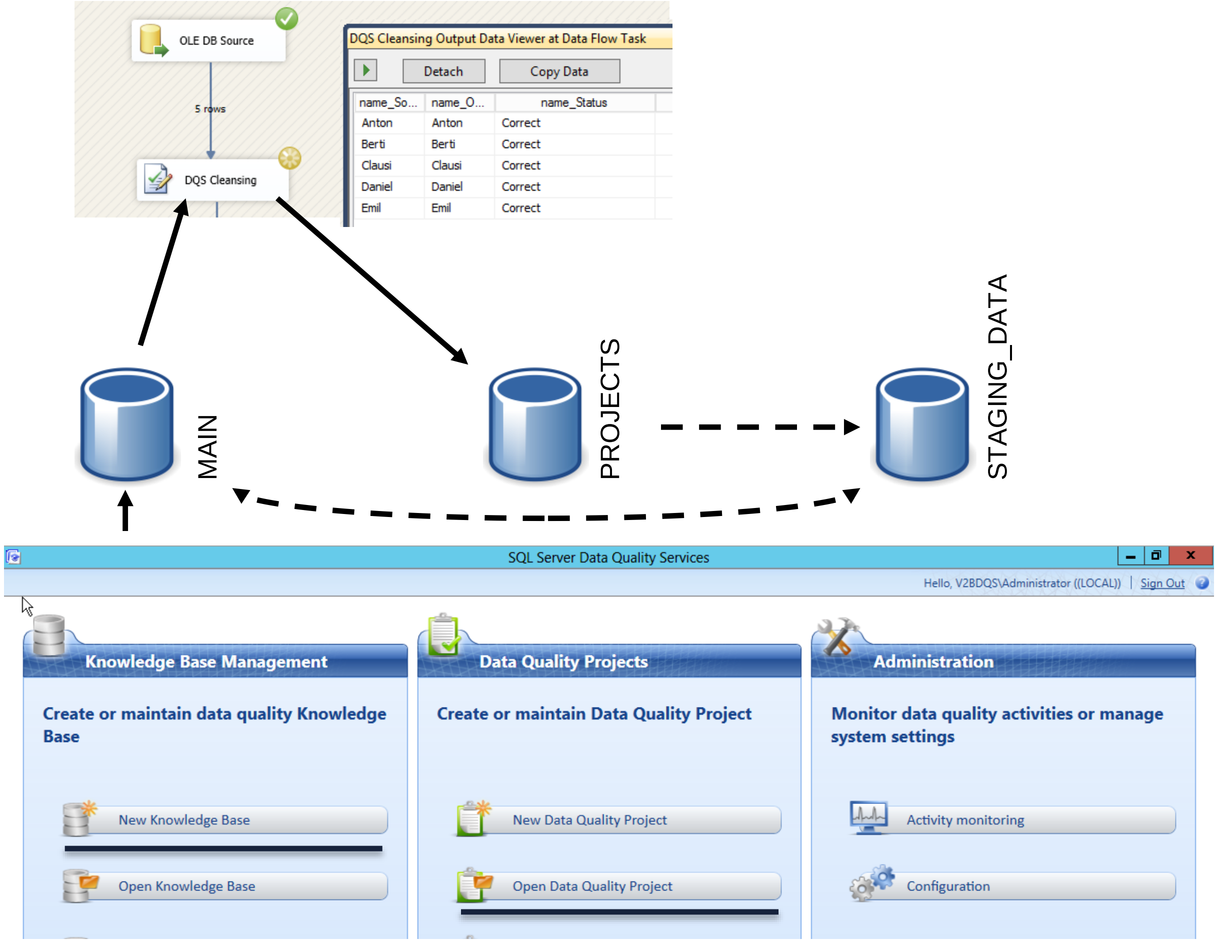
Data Quality Projects

Create or maintain Data Quality Project

Administration

Monitor data quality activities or manage system settings

DQS Übersicht



DQS Cleansing Output Data Viewer at Data Flow Task

name_So...	name_O...	name_Status
Anton	Anton	Correct
Berti	Berti	Correct
Clausi	Clausi	Correct
Daniel	Daniel	Correct
Emil	Emil	Correct

Create or maintain data quality Knowledge Base

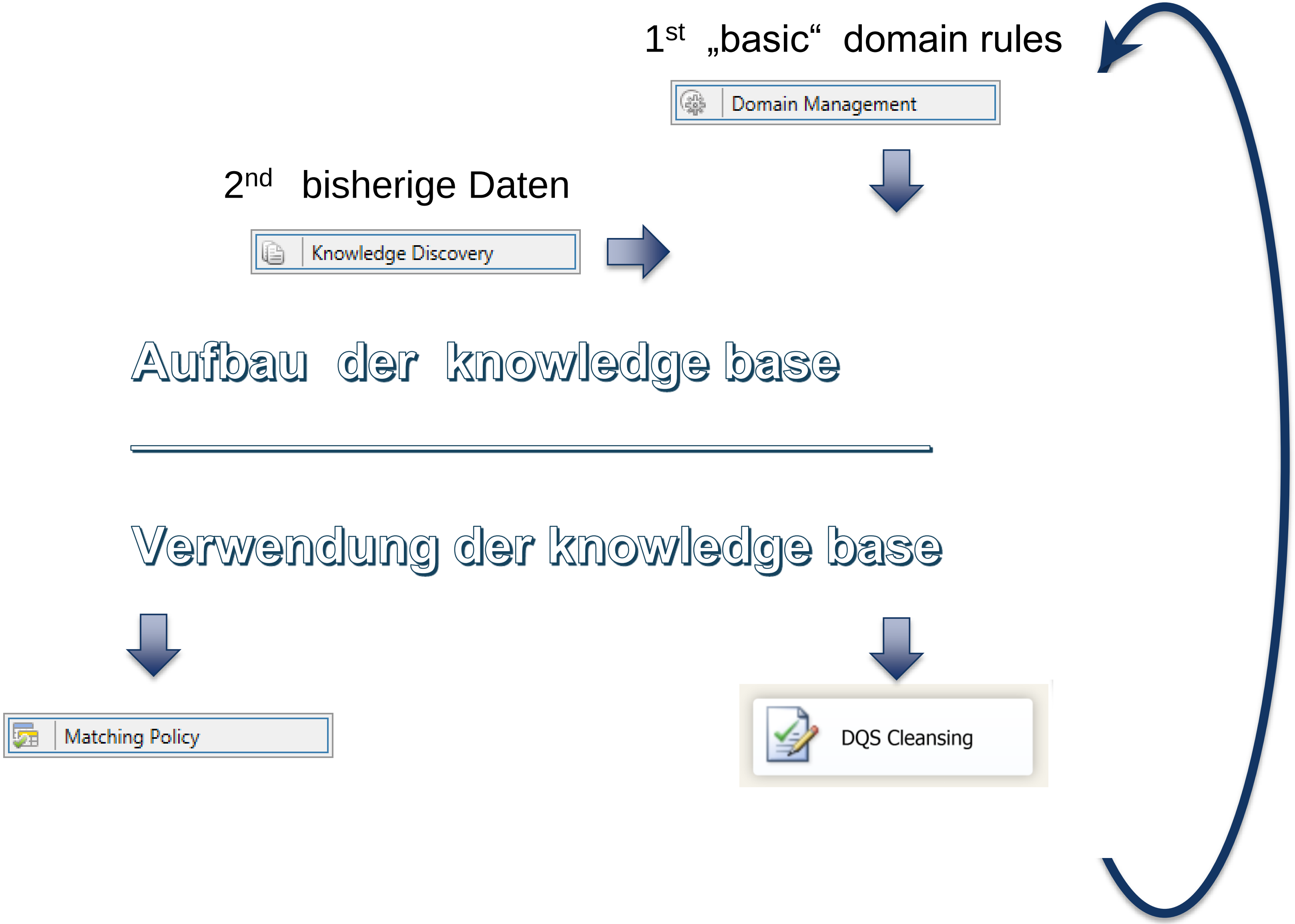
Create or maintain Data Quality Project

Monitor data quality activities or manage system settings

DQS zum Einstieg einige Begriffe

- Knowledgebase

- Domain Management
- Knowledge Discovery
- Matching Policy



1st „basic“ domain rules

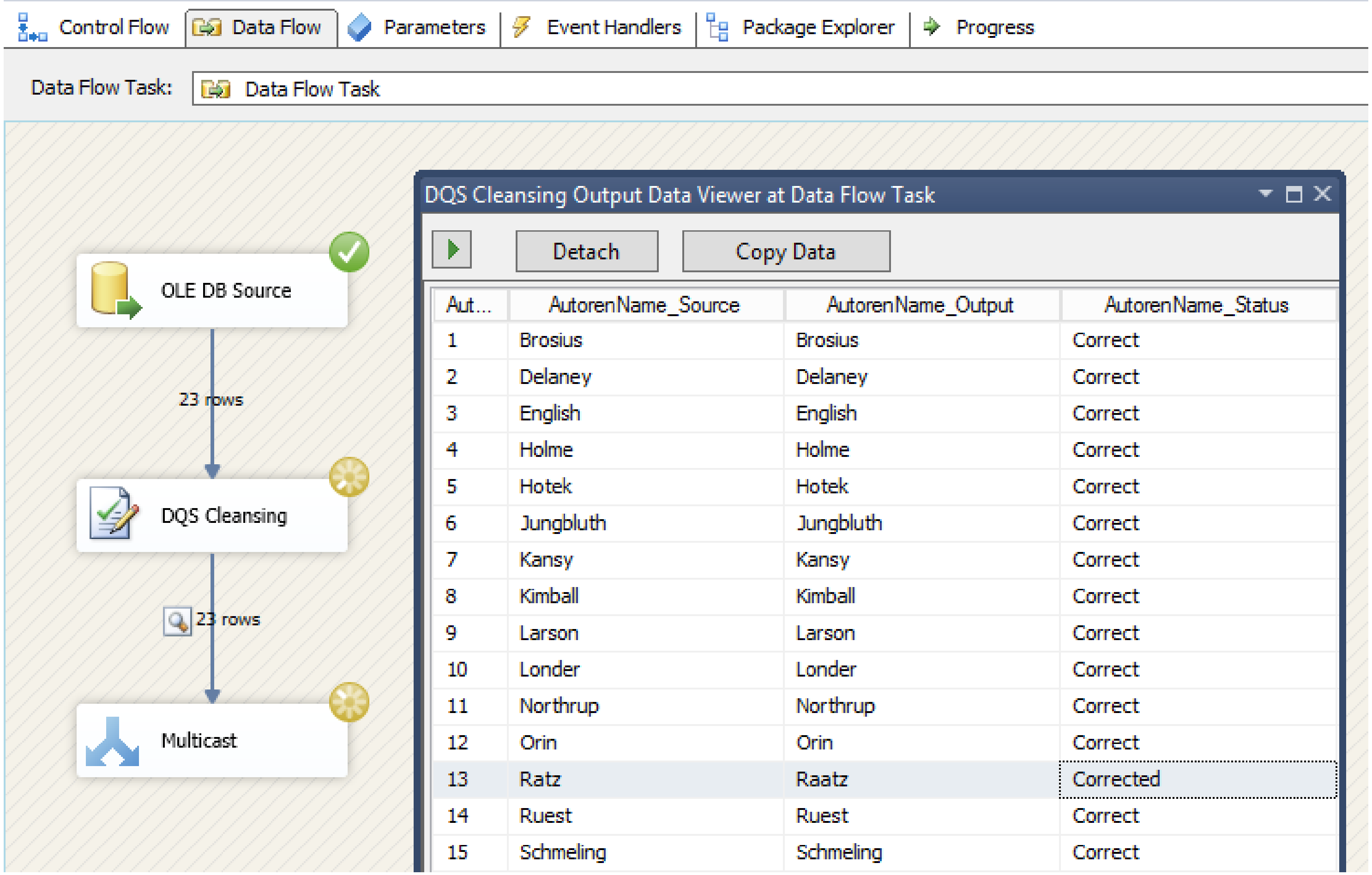
2nd bisherige Daten

Aufbau der knowledge base

Verwendung der knowledge base

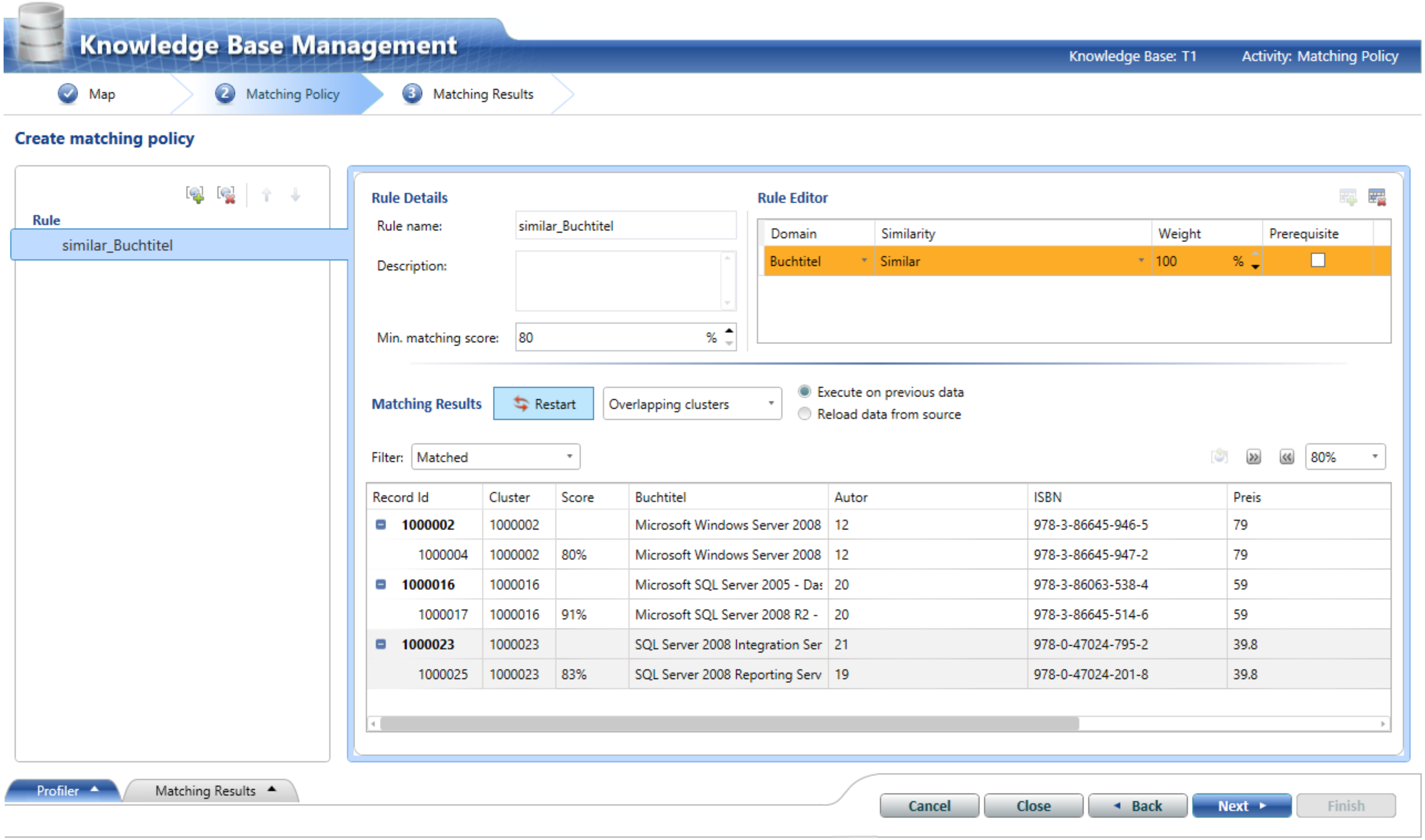
DEMO

Demo Results



Aut...	AutorenName_Source	AutorenName_Output	AutorenName_Status
1	Brosius	Brosius	Correct
2	Delaney	Delaney	Correct
3	English	English	Correct
4	Holme	Holme	Correct
5	Hotek	Hotek	Correct
6	Jungbluth	Jungbluth	Correct
7	Kansy	Kansy	Correct
8	Kimball	Kimball	Correct
9	Larson	Larson	Correct
10	Londer	Londer	Correct
11	Northrup	Northrup	Correct
12	Orin	Orin	Correct
13	Ratz	Raatz	Corrected
14	Ruest	Ruest	Correct
15	Schmeling	Schmeling	Correct

Demo 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

Overlapping clusters

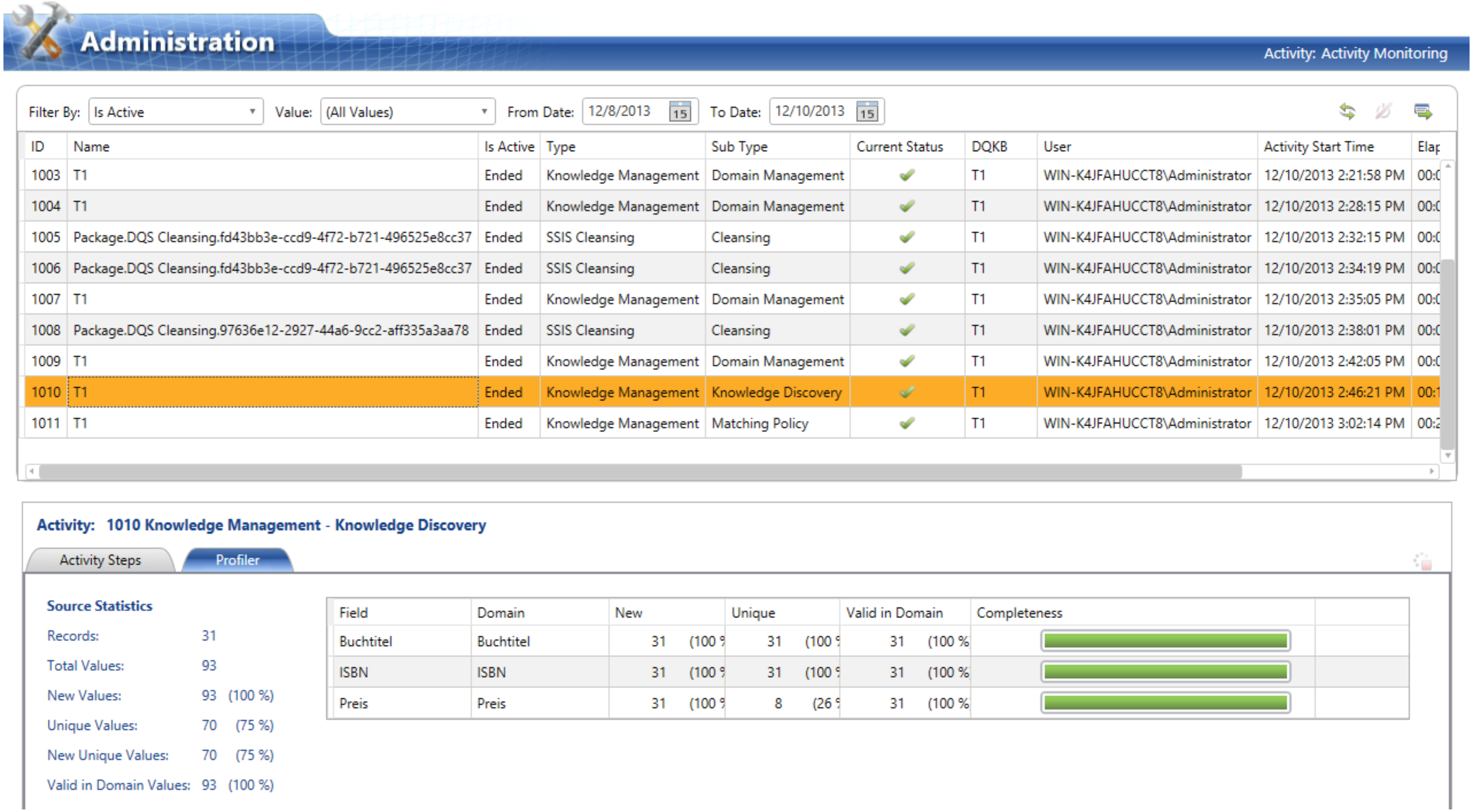
● Execute on previous data
○ Reload data from source

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

Demo Results



Filter By: Is Active Value: (All Values) From Date: 12/8/2013 To Date: 12/10/2013

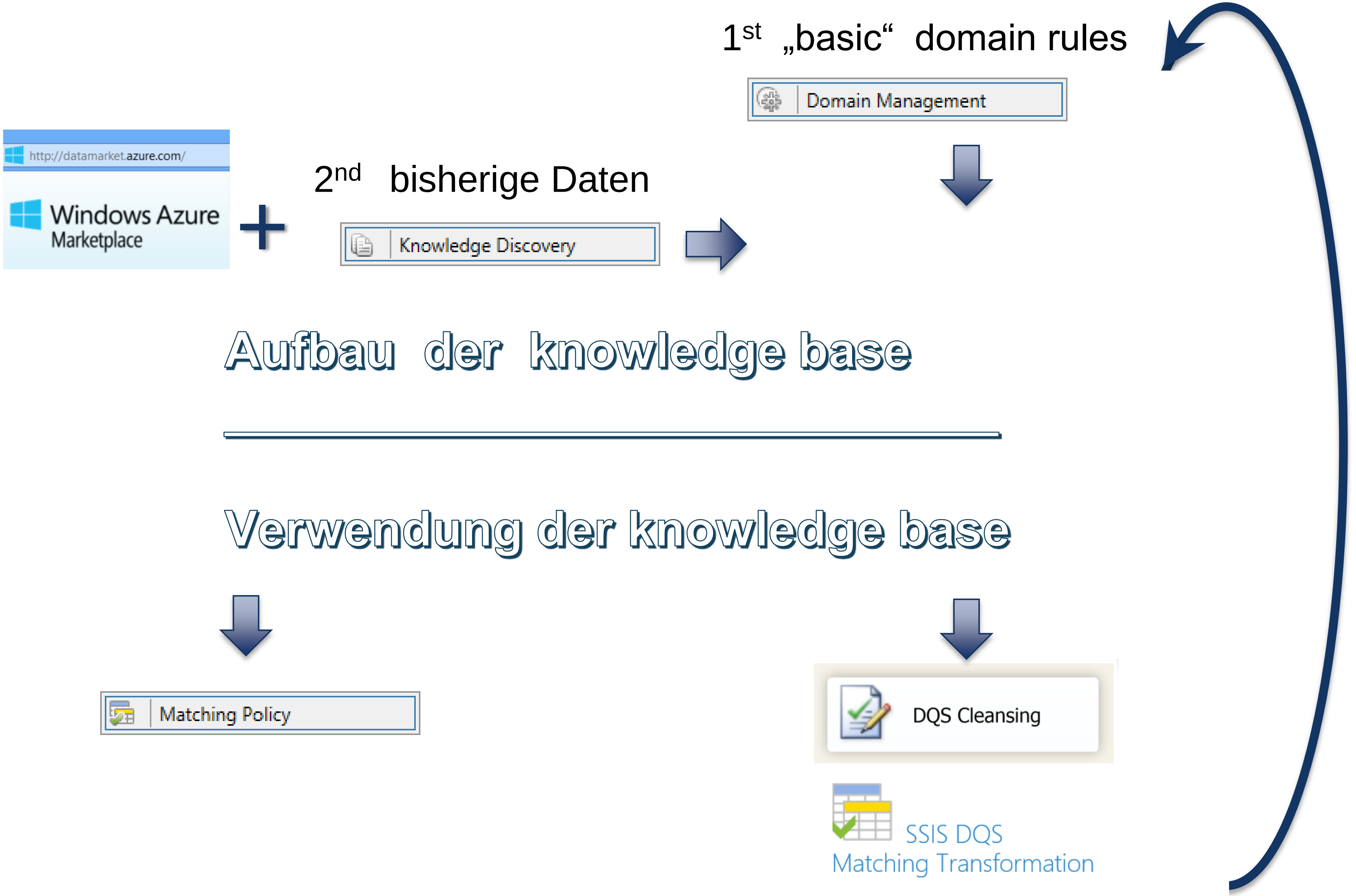
ID	Name	Is Active	Type	Sub Type	Current Status	DQKB	User	Activity Start Time	Elap
1003	T1	Ended	Knowledge Management	Domain Management	✓	T1	WIN-K4JFAHUCCT8\Administrator	12/10/2013 2:21:58 PM	00:0
1004	T1	Ended	Knowledge Management	Domain Management	✓	T1	WIN-K4JFAHUCCT8\Administrator	12/10/2013 2:28:15 PM	00:0
1005	Package.DQS Cleansing.fd43bb3e-ccd9-4f72-b721-496525e8cc37	Ended	SSIS Cleansing	Cleansing	✓	T1	WIN-K4JFAHUCCT8\Administrator	12/10/2013 2:32:15 PM	00:0
1006	Package.DQS Cleansing.fd43bb3e-ccd9-4f72-b721-496525e8cc37	Ended	SSIS Cleansing	Cleansing	✓	T1	WIN-K4JFAHUCCT8\Administrator	12/10/2013 2:34:19 PM	00:0
1007	T1	Ended	Knowledge Management	Domain Management	✓	T1	WIN-K4JFAHUCCT8\Administrator	12/10/2013 2:35:05 PM	00:0
1008	Package.DQS Cleansing.97636e12-2927-44a6-9cc2-aff335a3aa78	Ended	SSIS Cleansing	Cleansing	✓	T1	WIN-K4JFAHUCCT8\Administrator	12/10/2013 2:38:01 PM	00:0
1009	T1	Ended	Knowledge Management	Domain Management	✓	T1	WIN-K4JFAHUCCT8\Administrator	12/10/2013 2:42:05 PM	00:0
1010	T1	Ended	Knowledge Management	Knowledge Discovery	✓	T1	WIN-K4JFAHUCCT8\Administrator	12/10/2013 2:46:21 PM	00:1
1011	T1	Ended	Knowledge Management	Matching Policy	✓	T1	WIN-K4JFAHUCCT8\Administrator	12/10/2013 3:02:14 PM	00:2

Activity: 1010 Knowledge Management - Knowledge Discovery

Source Statistics

Records:	31	
Total Values:	93	
New Values:	93	(100 %)
Unique Values:	70	(75 %)
New Unique Values:	70	(75 %)
Valid in Domain Values:	93	(100 %)

Field	Domain	New		Unique		Valid in Domain		Completeness
Buchtitel	Buchtitel	31	(100	31	(100	31	(100 %	🟩
ISBN	ISBN	31	(100	31	(100	31	(100 %	🟩
Preis	Preis	31	(100	8	(26	31	(100 %	🟩



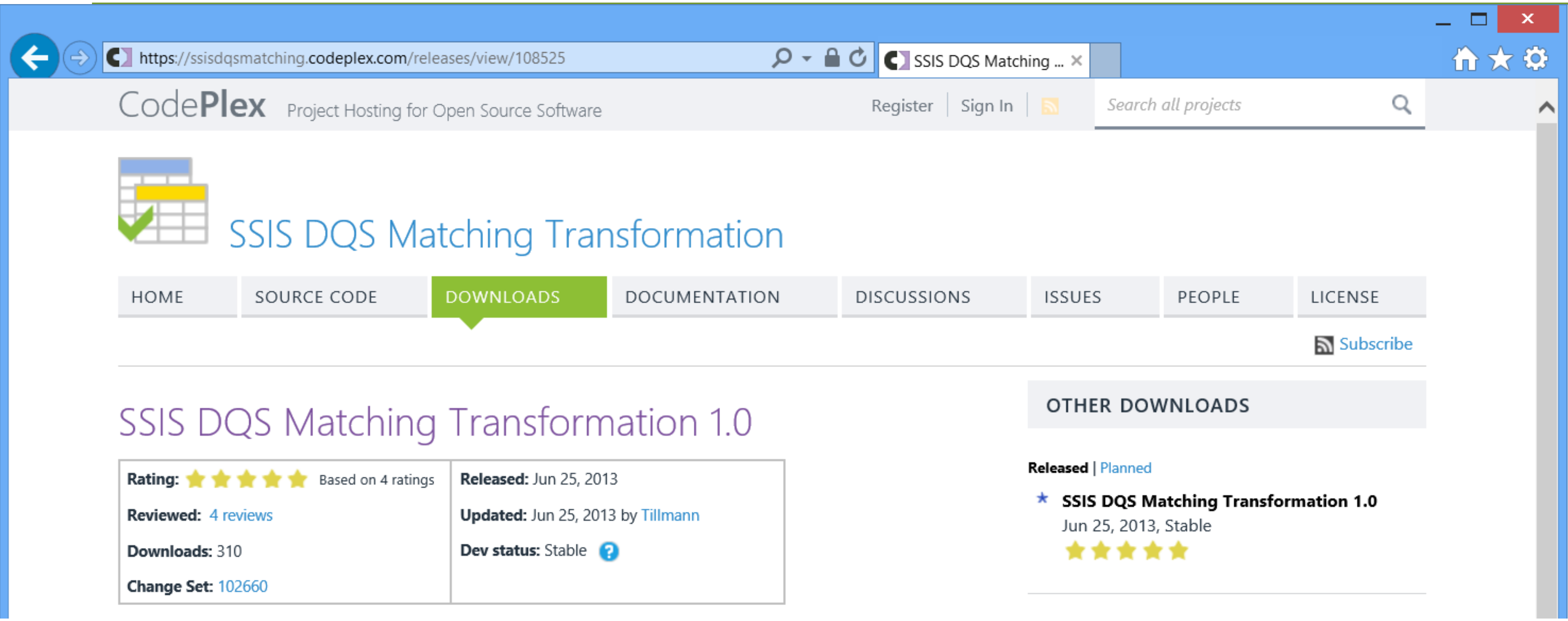
1st „basic“ domain rules

2nd bisherige Daten

Aufbau der knowledge base

Verwendung der knowledge base

Domain Management
http://datamarket.azure.com/
Windows Azure Marketplace
+
Knowledge Discovery
Matching Policy
DQS Cleansing
SSIS DQS Matching Transformation



SSIS DQS Matching Transformation

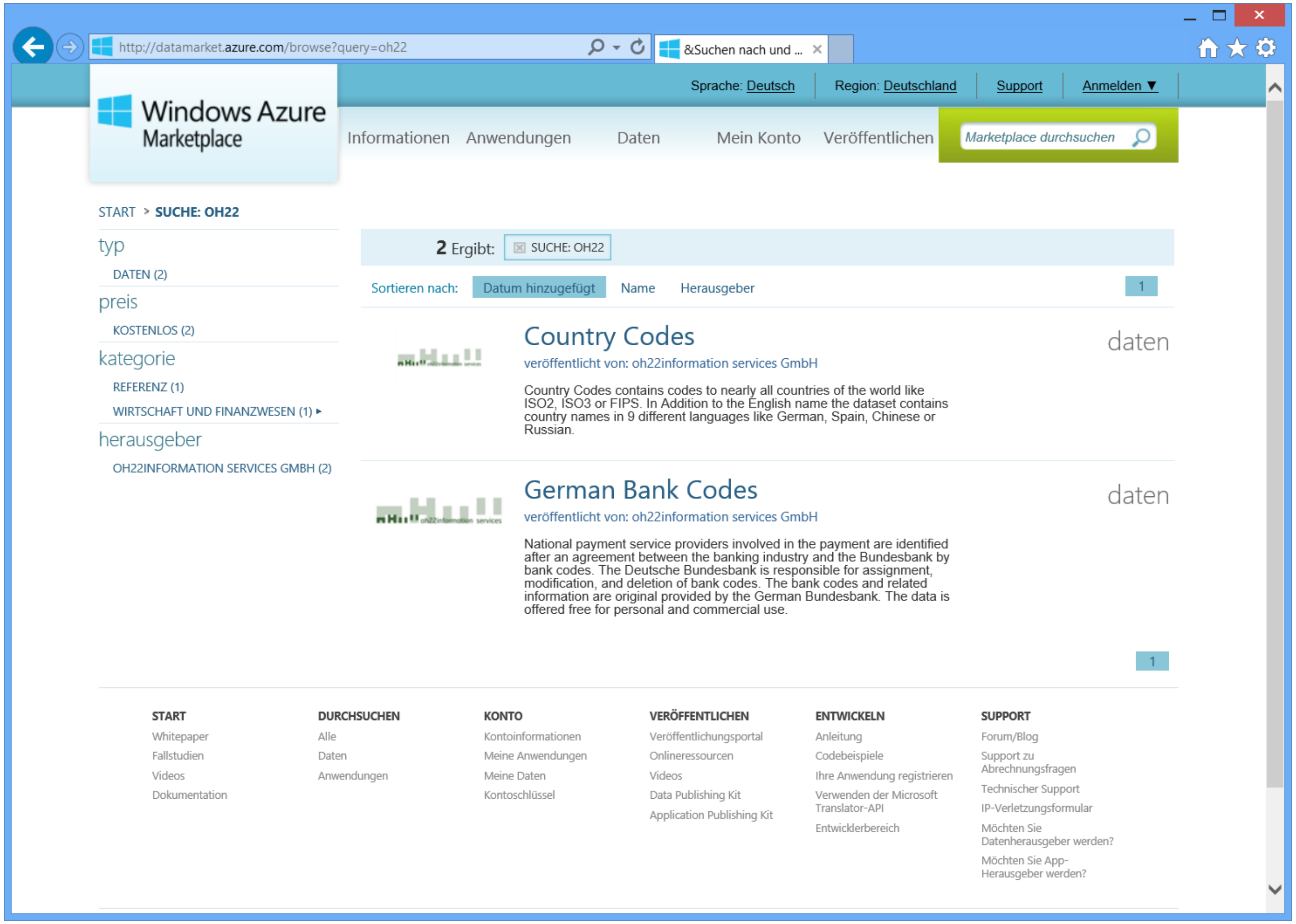
SSIS DQS Matching Transformation 1.0

Rating: ★★★★★ Based on 4 ratings	**Released:** Jun 25, 2013
Reviewed: 4 reviews	**Updated:** Jun 25, 2013 by Tillmann
Downloads: 310	**Dev status:** Stable
Change Set: 102660	

OTHER DOWNLOADS

Released | Planned

★ **SSIS DQS Matching Transformation 1.0**
Jun 25, 2013, Stable
★★★★★



Country Codes

veröffentlicht von: oh22information services GmbH

daten

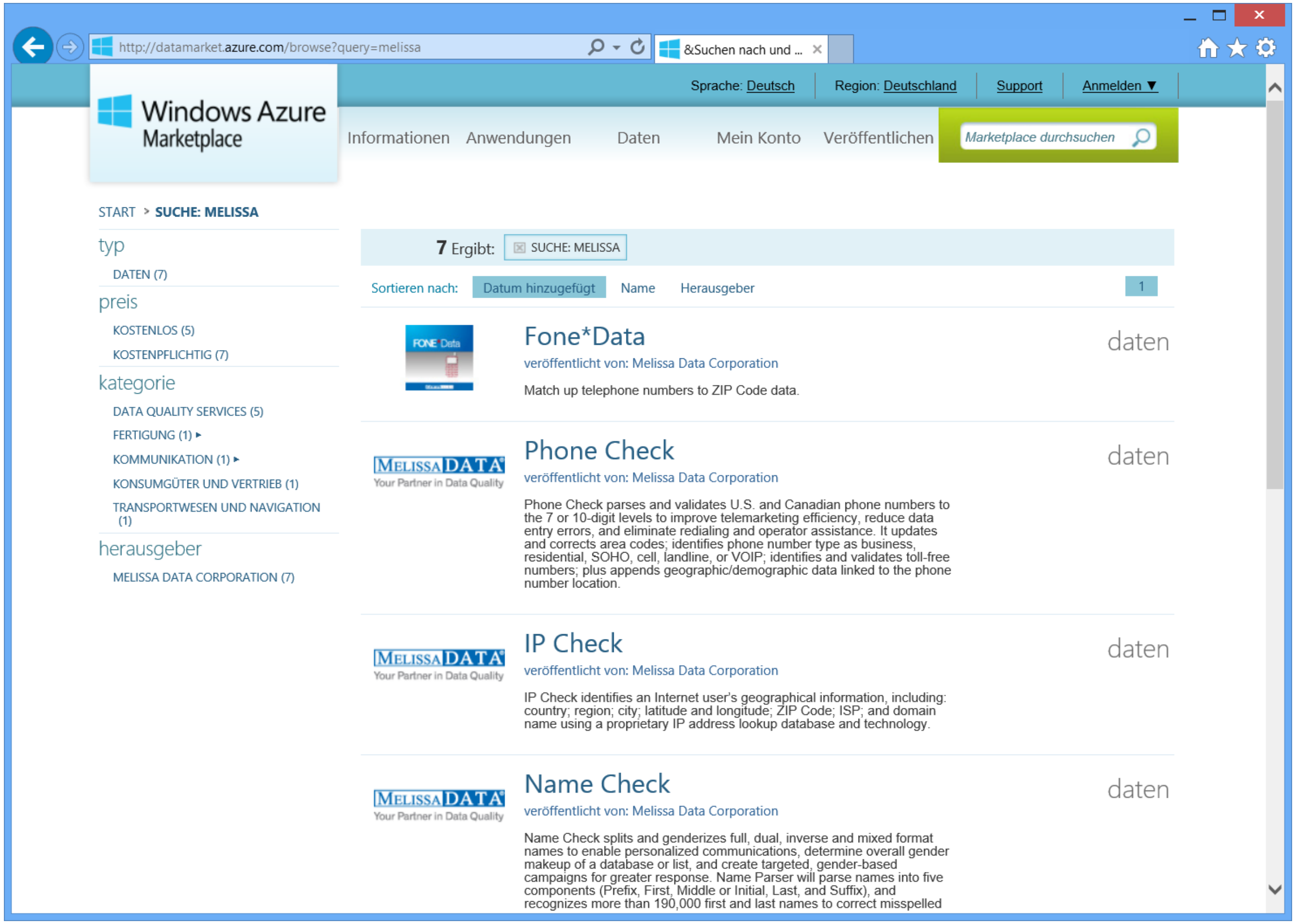
Country Codes contains codes to nearly all countries of the world like ISO2, ISO3 or FIPS. In Addition to the English name the dataset contains country names in 9 different languages like German, Spain, Chinese or Russian.

German Bank Codes

veröffentlicht von: oh22information services GmbH

daten

National payment service providers involved in the payment are identified after an agreement between the banking industry and the Bundesbank by bank codes. The Deutsche Bundesbank is responsible for assignment, modification, and deletion of bank codes. The bank codes and related information are original provided by the German Bundesbank. The data is offered free for personal and commercial use.



START > **SUCHE: MELISSA**

typ
- DATEN (7)

preis
- KOSTENLOS (5)
- KOSTENPFLICHTIG (7)

kategorie
- DATA QUALITY SERVICES (5)
- FERTIGUNG (1) ▸
- KOMMUNIKATION (1) ▸
- KONSUMGÜTER UND VERTRIEB (1)
- TRANSPORTWESEN UND NAVIGATION (1)

herausgeber
- MELISSA DATA CORPORATION (7)

7 Ergibt: ☒ SUCHE: MELISSA

Sortieren nach: Datum hinzugefügt | Name | Herausgeber — 1

Fone*Data — daten

veröffentlicht von: Melissa Data Corporation

Match up telephone numbers to ZIP Code data.

Phone Check — daten

veröffentlicht von: Melissa Data Corporation

Phone Check parses and validates U.S. and Canadian phone numbers to the 7 or 10-digit levels to improve telemarketing efficiency, reduce data entry errors, and eliminate redialing and operator assistance. It updates and corrects area codes; identifies phone number type as business, residential, SOHO, cell, landline, or VOIP; identifies and validates toll-free numbers; plus appends geographic/demographic data linked to the phone number location.

IP Check — daten

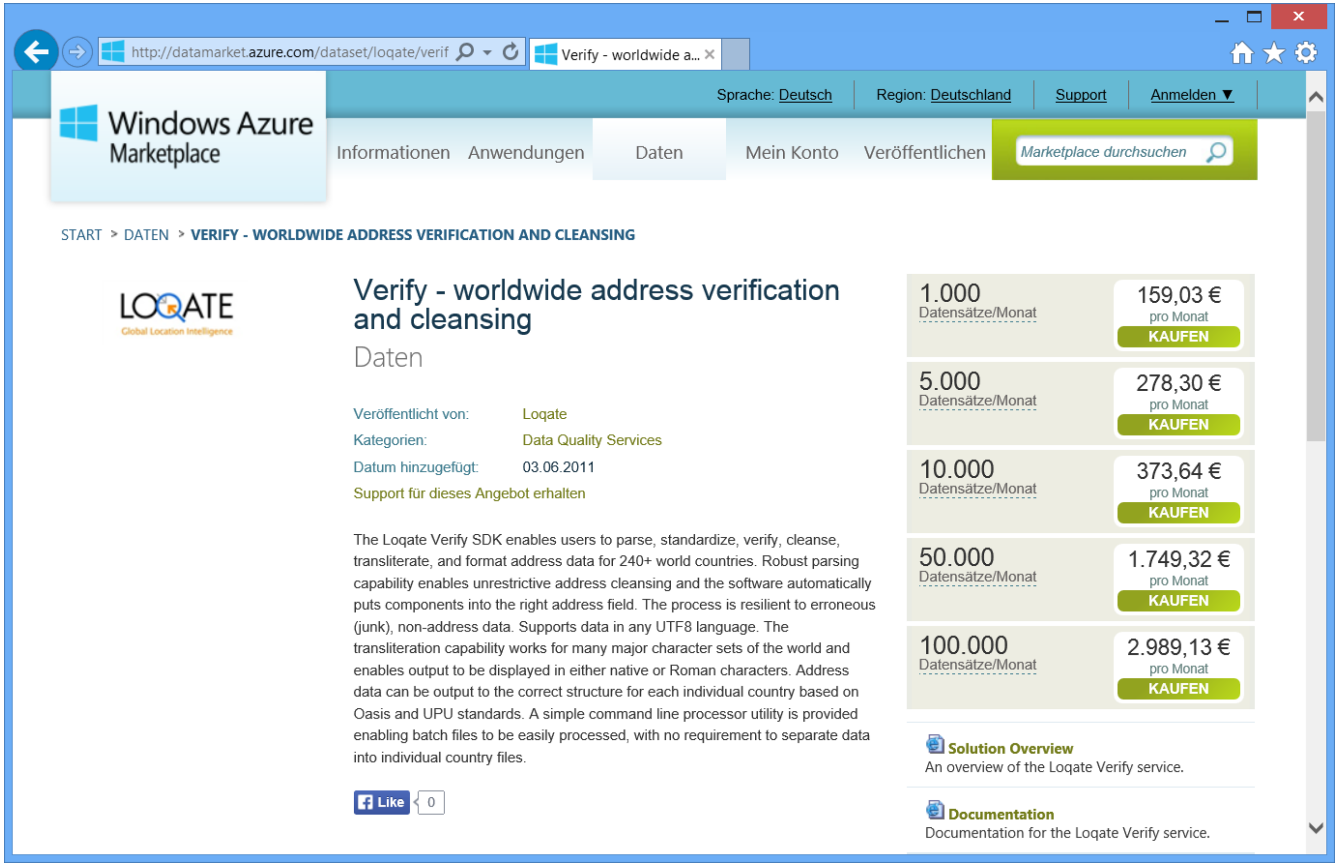
veröffentlicht von: Melissa Data Corporation

IP Check identifies an Internet user's geographical information, including: country; region; city; latitude and longitude; ZIP Code; ISP; and domain name using a proprietary IP address lookup database and technology.

Name Check — daten

veröffentlicht von: Melissa Data Corporation

Name Check splits and genderizes full, dual, inverse and mixed format names to enable personalized communications, determine overall gender makeup of a database or list, and create targeted, gender-based campaigns for greater response. Name Parser will parse names into five components (Prefix, First, Middle or Initial, Last, and Suffix), and recognizes more than 190,000 first and last names to correct misspelled


http://datamarket.azure.com/dataset/loqate/verif
Verify - worldwide a...
Sprache: Deutsch
Region: Deutschland
Support
Anmelden ▼
Windows Azure Marketplace
Informationen
Anwendungen
Daten
Mein Konto
Veröffentlichen
Marketplace durchsuchen
START > DATEN > VERIFY - WORLDWIDE ADDRESS VERIFICATION AND CLEANSING
LOQATE
Global Location Intelligence

Verify - worldwide address verification and cleansing

Daten

Veröffentlicht von:	Loqate
Kategorien:	Data Quality Services
Datum hinzugefügt:	03.06.2011

Support für dieses Angebot erhalten

The Loqate Verify SDK enables users to parse, standardize, verify, cleanse, transliterate, and format address data for 240+ world countries. Robust parsing capability enables unrestrictive address cleansing and the software automatically puts components into the right address field. The process is resilient to erroneous (junk), non-address data. Supports data in any UTF8 language. The transliteration capability works for many major character sets of the world and enables output to be displayed in either native or Roman characters. Address data can be output to the correct structure for each individual country based on Oasis and UPU standards. A simple command line processor utility is provided enabling batch files to be easily processed, with no requirement to separate data into individual country files.

1.000 Datensätze/Monat	159,03 € pro Monat KAUFEN
5.000 Datensätze/Monat	278,30 € pro Monat KAUFEN
10.000 Datensätze/Monat	373,64 € pro Monat KAUFEN
50.000 Datensätze/Monat	1.749,32 € pro Monat KAUFEN
100.000 Datensätze/Monat	2.989,13 € pro Monat KAUFEN

Solution Overview
An overview of the Loqate Verify service.

Documentation
Documentation for the Loqate Verify service.

DQS Lösung

~~Nachteil „handmade“~~	standardisiert
~~ggf. Portierungsaufwand~~	gut portierbar
~~umfangreichere Doku~~	Doku vorhanden
~~Benutzer-Einweisung in workflow~~	nachvollziehbarer Workflow für Fachabt.

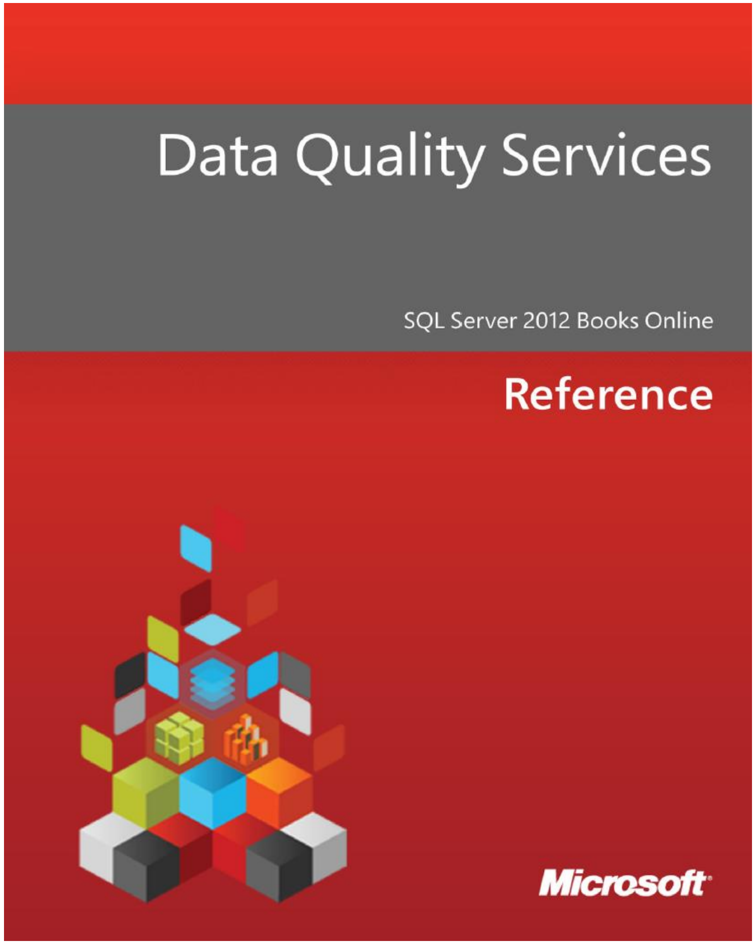
DQS Literaturempfehlung

download

DQS step-by-step

german / english

DQS Literaturempfehlung



Data Quality Services

SQL Server 2012 Books Online

Reference

Microsoft

vielen Dank