Dokumentation von BI Projekten

Dr. Yvette Teiken

21. Januar 2014

1 Einleitung

Dieses Dokument beschreibt wie BI Projekte durchgeführt werden sollen. Dieses Dokument stellt Grundlagen für die Dokumentation bereit.

2 Datenquellen

2.1 DataProfilingTaskSample

Name der Datenquelle: DataProfilingTaskSample **Format:** csv
Kurszbeschreibung: Ein einfache Datenquelle, die zum Testen gebaut habe

Attributname: ID **Beschreibung**: Fortlaufende ID
Attributname: Geschlecht **Beschreibung**: In der Kodierung m und w
Attributname: Alter **Beschreibung**: Alter als numerischer String. Werte sollen zwischen 0 und 130 liegen
Attributname: Vermögen **Beschreibung**: Numerischer Ganzzahl String

3 Datenintegration

3.1 DataProfilingTaskSample

3.2 Extraktion

Hier steht der beschreibne Text, wie das realisiert
Bsonderheiten: Hier musste man eine Ausnahme von dem Vorgen in 9.2 gemacht, da hier die Daten anders sind.

3.3 Kodierung

3.4 Struktur

3.5 Umsetzung Fachlichkeit

Hier muss die Regel 9.2 Anwendung finden, da ... Realisiert wurde dies mit der SIS Komponenten.

4 Datenintegration: DataProfilingTaskSample

Ein einfaches Beispiel für die Datenquellen

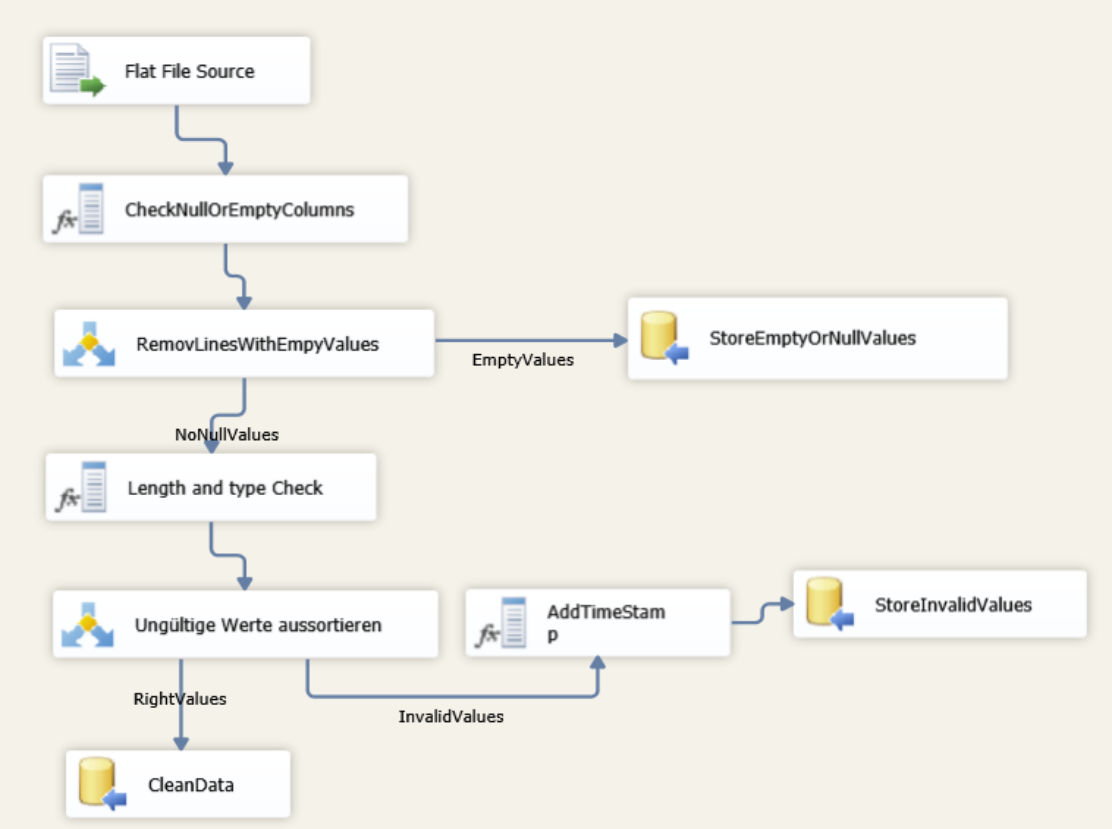


Abbildung 1: Vollständigkeit und Struktur SISS Paket

4.1 Extraktion

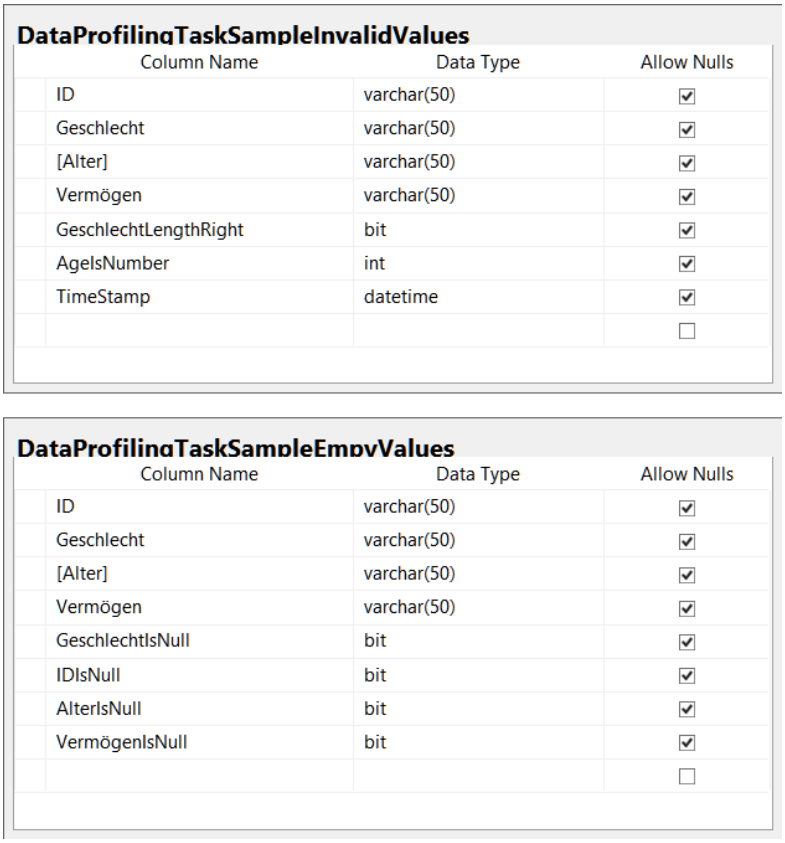
Hier wird mittels csv flatfile Import realisiert **Bsonderheiten:** Hier musste man eine Ausnahme von dem Vorgen in 9.2 gemacht, da hier die Daten anders sind.

4.2 Kodierung

4.3 Vollständigkeit und Struktur

Die Vollständigkeit und die Struktur wird in dem Paket *Bereinigung.dtsx* realisiert. Und ist in der Abbildung dargestellt. Zuerst wird auf leere oder null-Werte überprüft und aussortiert. Dies passiert im Task *CheckNullOrEmptyColumns* . Haben Zeilen leere Werte, werden diese komplett aussortiert und in die Tabelle eingefügt

Die Länge der Datentypen und die Formatieren werden im Task *Length and type Check* ausgeführt. Hier wird die Länge des Geschlecht und ob es sich bei Alter um eine Zahl handelt.



DataProfilingTaskSampleInvalidValues

Column Name	Data Type	Allow Nulls
ID	varchar(50)	☑
Geschlecht	varchar(50)	☑
[Alter]	varchar(50)	☑
Vermögen	varchar(50)	☑
GeschlechtLengthRight	bit	☑
AgeIsNumber	int	☑
TimeStamp	datetime	☑
		☐

DataProfilingTaskSampleEmpyValues

Column Name	Data Type	Allow Nulls
ID	varchar(50)	☑
Geschlecht	varchar(50)	☑
[Alter]	varchar(50)	☑
Vermögen	varchar(50)	☑
GeschlechtIsNull	bit	☑
IDIsNull	bit	☑
AlterIsNull	bit	☑
VermögenIsNull	bit	☑
		☐

Abbildung 2: Logging Tabellen

SISS Paket

Verwendete Logging Tabellen Überprüft werden müssen die Felder

4.4 Umsetzung Fachlichkeit

Hier muss die Regel 8.2 Anwendung finden, da ... Realisiert wurde dies mit der SSIS Komponente *Fachliche Regel Alter*

```
(DT_I4)Alter >= 0 && (DT_I4)Alter <= 130
```

Realisiert in Package: ExtractIncidentReport.dtsx

5 Zentrale Datenhaltung

6 Aggregation

7 Reporting

8 Fachliche Reglen

8.1 Geschlecht zu ICD

8.2 Wertebereich Alter:

Ein Alter darf nur zwischen 0 und 130 liegen

Fachliche Beschreibung: wenn ein ICD 40 oder 50 lautet dann kann das Geschlecht nur weiblich sein.

Domänenexperte Herr Sander

9 Architektur- und Programmierrichtlinien

In diesem Abschnitt werden Richtlinien und Best Practices gesammelt.

9.1 Formatierungen

Es gibt zwei Alternativen hierfür:

Mittels SSIS: Die Konvertierungen werden mittels SISS Komponenten durchgeführt.Wie dies geht, hier **??** vorgestellt. Hier verwendet man abgeleitete Spalten für

Listing 1: Formatierung mit SISS und Behandlung von leeren Werten

```
% reine Prüfung
(DT_UI4)Alter == (DT_UI4)Alter ? 1 : 0
% hier wird es automatisch auf den konvertireten Wert gelegt
(DT_UI4)Alter == (DT_UI4)Alter ? (DT_UI4)Alter : 0
```

Mittels SQL: Die Daten werden als Text in eine DB integriert und dann danach mittels SQL Konvertierungen in das Zielformat überführt.

9.2 Extraktion von Daten

9.3 Daten bereinigen mittels Data Profiling Task

Um den Task verwenden zu können, müssen die Daten in einer ADO.NET Datenquelle vorhanden sein. Es hilft auch nur um die Daten einschätzen zu können. Die Ergebnisse liegen dann in einer XML Datei und können betrachtet werden. Man sieht in den Ergebnisse eine Verteilung von Werten.

9.3.1 Daten bereinigen mittels DQS

Der Ansatz an sich ist ganz cool. Mann kann Domänen festlegen und Wertebereiche und Regeln zu den Werten festlegen. Der Client an sich ist relativ langsam und etwas fehleranfällig und absturzgefährdet. Ein kontinuierliches Arbeiten mit dem Client ist fast unmöglich.

In der Abbildung 3 werden die möglichen Ausprägungen beschrieben. Es können Synonyme verwendet werden, diese werden dann auf den Repräsentanten abgebildet. Auch können Regeln festgelegt werden. Wie im Beispiel Alter zu sehen ist. Hier wird der Wertebereich entsprechen eingeschränkt.

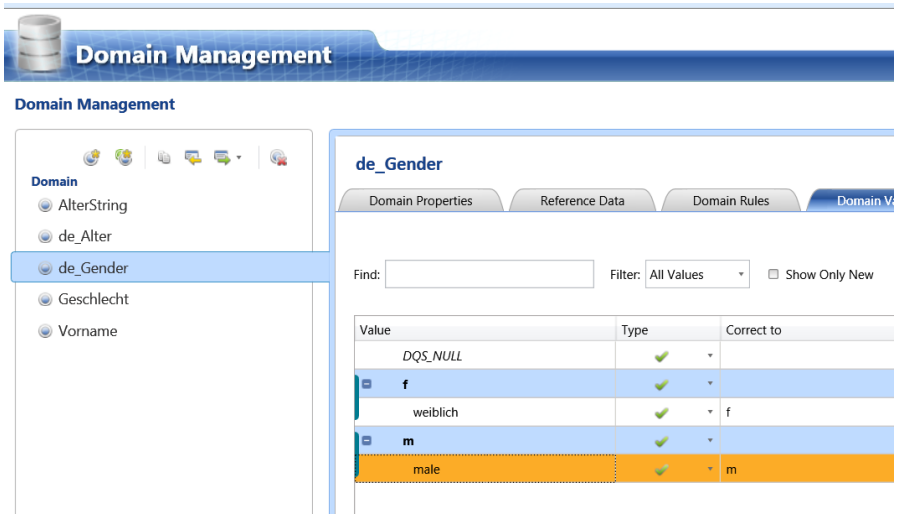


Abbildung 3: Domän Geschlecht mit Werten

Die Bereinigung der Daten kann dann im SISS vorgenommen werden, wie es in Abbildung 4 gezeigt ist. Der Task an sich ist recht schön. Man kann die Domäne zu einer Spalte angeben und alle Regelen, die in der Kowledgebase verfügbar sind werden angewendet. Als Ergebnis erhält man dann korrigierte oder aussortierte Daten, wie in Abbildung 5 zu sehen ist.[1]

[1] In meinem Beispiel hat die Abbildung von Geschlecht männlich nicht funktioniert und es wurde die ID der Spalte verwendet. Ich konnte nicht nachvollziehen warum das so ist.

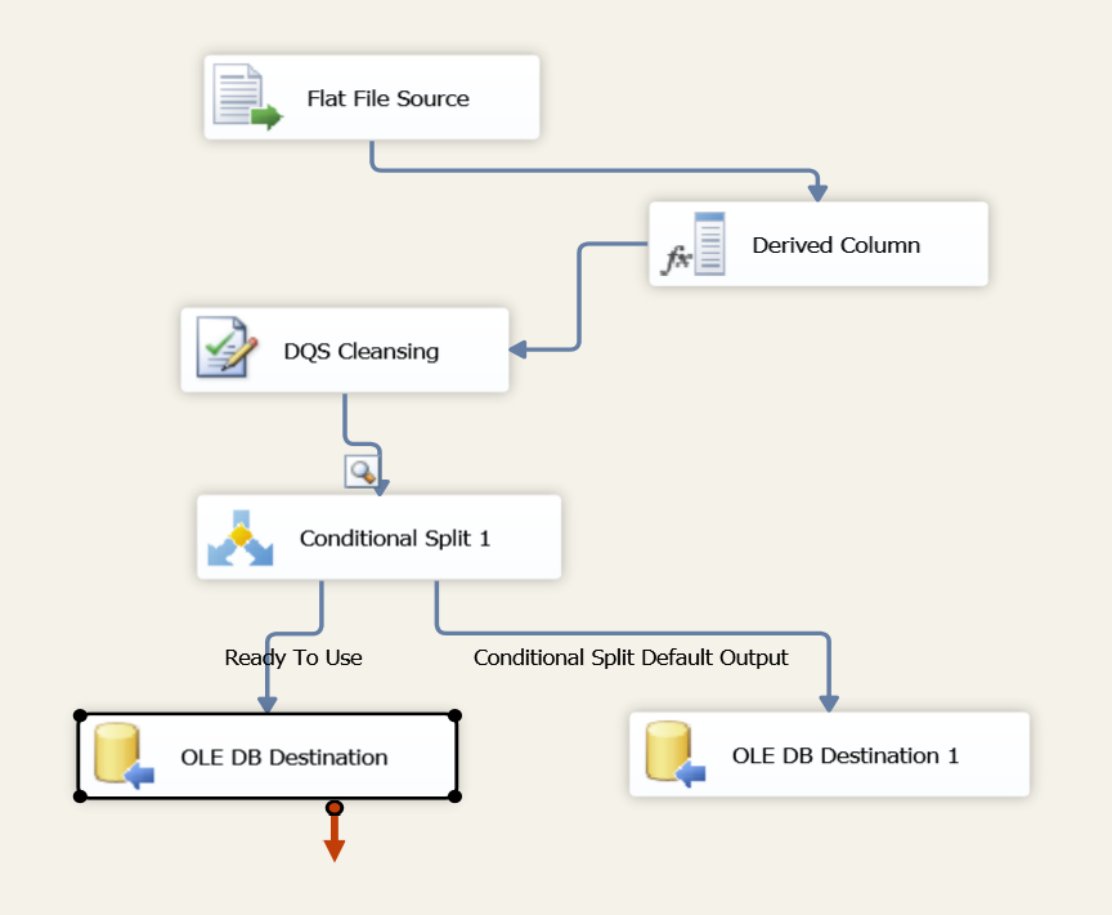


Abbildung 4: DQS Task in SISS

	ID	Geschlecht_Sou...	Geschlecht_Output	Geschlecht_Status	Alter_Source	Alter_Out...	Alter_Stat...	Vermög...	Record Status
1	2	m	2	Correct	1233	324	Invalid	324	Invalid

Abbildung 5: DQS Ergebins, invalide Daten

Index