Dokumentation von BI Projekten

Dr. Yvette Teiken

23. Januar 2014

Inhaltsverzeichnis

0.1 Einleitung

Dieses Dokument beschreibt wie BI Projekte durchgeführt werden sollen. Dieses Dokument stellt Grundlagen für die Dokumentation bereit. Hier werden Checklisten und Best Practices gesammelt und zusammengefügt.

Kapitel 1

Allgemeine Architektur

In diesem Kapitel wird die generelle Architektur der Lösung bzw. des Projekts beschrieben.

1.1 Datenanlieferung

Was ist das besondere an der Anlieferung, wo und wie muss das berücksichtigt werden

- □ Wie werden die Daten angeliefert?
- □ Werden alle Daten gleich angeliefert bzw. welche Ausnahmen gibt es?
- □ Gibt es besondere Anforderungen an Sicherheit bzw. Aufbewahrungspflichten?
- ☑ Das ist schon erledigt.

1.2 Abgeleitete Architektur

Wie sieht die Architektur aus? Welche Schichten gibt es? Wie ist der Datenfluss zwischen den Schichten? Wie ist der Bezug zu unseren Best Practices?

1.3 Logging und Bereinigung

- □ Wie soll das Logging laufen?
- □ Was sind die Prozesse?
- □ Welche Tabellen werden beteiligt?

1.4 Zentrale Datenhaltung

- □ Wie ist das zentrale Analyseformat (Star, Data Vault oder was anderes?)
- □ Wie wird mit Business Keys umgegangen?
- □ Wie werden Stammdaten abgeleitet?

1.5 Berichte

Welche Berichte gibt es? WIe werden diese erzeugt?

Kapitel 2

Konkrete Umsetzung des Projekts

Hier wird an Hand der konkreten Datenquellen bzw. Berichte die einzelnen Schichten beschrieben.

2.1 DataProfilingTaskSample

2.1.1 Datenquellenbeschreibung

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: VermÃ¶gen **Beschreibung**: Numerischer Ganzzahl String

2.1.2 Extraktion

Hier wird mittels csv flatfile Import realisiert

2.1.3 Kodierung

2.1.4 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

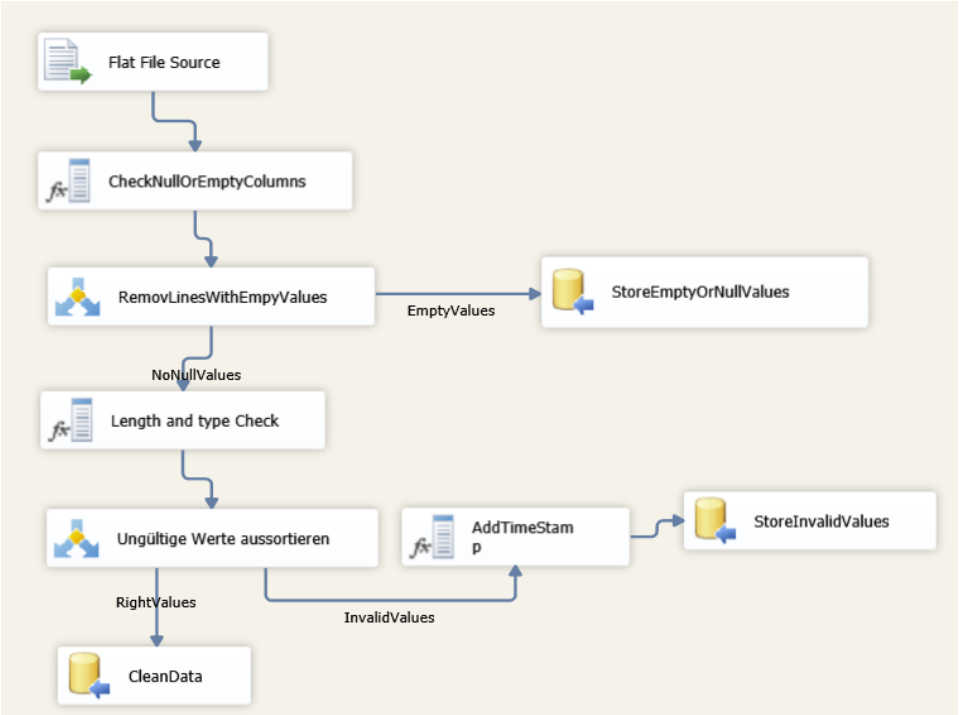


Abbildung 2.1: Vollständigkeit und Struktur SISS Paket

Die Länge der Datentypen und die Formatieren werden im Task *Length and type Check* ausgefÃ¼hrt. Hier wird die Länge des Geschlecht und ob es sich bei Alter um eine Zahl handelt.

SISS Paket

Verwendete Logging Tabellen

2.1.5 Umsetzung Fachlichkeit

Hier muss die Regel 3.2 Anwendung finden, da ... Realisiert wurde dies mit der SSIS Komponente *Fachliche Regel Alter*

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

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 2.2: Logging Tabellen

Kapitel 3

Fachliche Regeln

3.1 Geschlecht zu ICD

3.2 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.

Kapitel 4

Architektur- und Programmierrichtlinien

In diesem Abschnitt werden Richtlinien und Best Practices gesammelt.

4.1 Technichnical IDs

Für die Vergabe von technischen Schlüsseln gibt es zwei Alternativen als numeric oder als Guid

Eigenschaften von GUIDs

- Sind global eindeutig im System, ein Item ist immer eindeutig identifizierbar
- neue Guids im SSIS kann man mittels abgeleiteter Spalte hinzufügen zum Datenfluss
- Sind in der Geschwindigkeit bei der Verarbeitung gefühlt langsamer
- Bei der DAK und im Data Vault Kontext werden Guids für technische IDs verwendet.

Eigenschaften von numerischen technischen Schlüsseln

- Können automatisch erzeugt werden beim Einfügen in der Tabelle
- Sind nur eindeutig pro Tabelle und nicht im System
- Sind in der Verarbeitung etwas angenehmer

4.2 Stagging in Tabellen

4.3 Formatierungen

Es gibt zwei Alternativen hierfür:

4.3.1 Mittels SSIS:

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

Listing 4.1: 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.

4.3.2 Extraktion von Daten

4.3.3 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.

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 4.1 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.

Die Bereinigung der Daten kann dann im SISS vorgenommen werden, wie es in Abbildung 4.2 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 4.3 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.

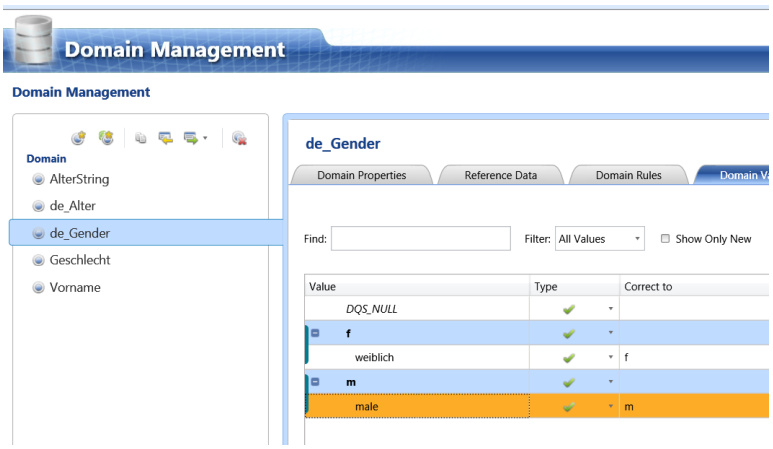


Abbildung 4.1: Domän Geschlecht mit Werten

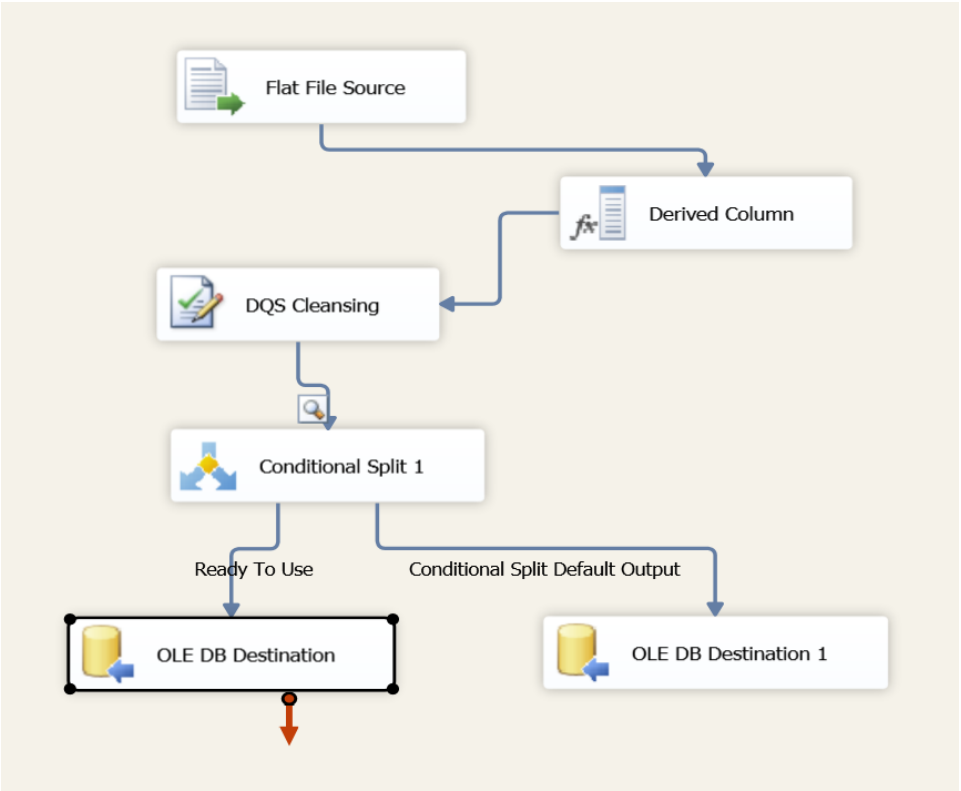
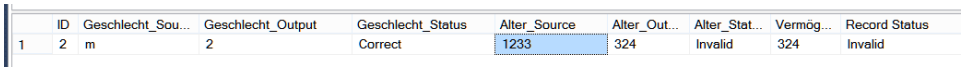


Abbildung 4.2: DQS Task in SISS



	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 4.3: DQS Ergebins, invalide Daten

References

[Dus] DustinRyan. Check isnumeric() with derived column transform in ssis package.

Index