Automatische Korrektur der NULL-Werte durch Defaultwerte (Singlestones)

Feraydoun Mohajeri
Triestram & Partner
Bochum

Schlüsselworte:
Defaultwerte, Null-Handling, Singlestones, BI, DWH, Early Arriving Facts, Late Arriving Dimensions

Einleitung
In den vielen BI-Tools werden bei der Erstelllung der Reports automatische SQL-Statements generiert, in denen bei der Verknüpfung der Dimension- und Faktentabellen immer ein „INNER JOIN“ benutzt wird.

Falls aber die ID der Dimensionen in der Faktentabelle zum Zeitpunkt des Ladens noch nicht geliefert werden kann, wird der entsprechende Datensatz im Report nicht mehr angezeigt. Das ist ein typisches Problem beim Laden der Daten in einem DWH.

Zur Lösung dieser sog. „Early Arriving Facts“ oder „Late Arriving Dimensions“ gibt es unterschiedliche Ansätze, die bei den unterschiedlichen Projekten zum Einsatz kommen und über die mehrfach diskutiert worden ist.

Hier wird nur auf die fehlenden Referenzen auf Dimensionen und Fakten eingegangen, die durch Defaultwerte (Singeltones) zu ersetzen sind. Die Problematik auf Grund der unbekannten Attribute oder doppelten Sätze werden hier nicht behandelt.

Automatische Korrektur der Null-Werte
Es soll mit Hilfe von dynamischem SQL ein Self-Cleaning-Mechanismus aufgebaut werden, mit dessen Hilfe die fehlenden Attribute automatisch vor dem Einfügen in die Core-Datenbank korrigiert werden.

Die folgenden Beispiele sollen die Problematik besser veranschaulichen.

Dimensionstabellen	Faktentabellen
D_BLOCKING	F_SALES
D_SUPPL_TYPE	F_ORDERS
D_ORDER_TYPE	F_RECEIPTS
D_COUNTRY	
D_REGION	
D_SEASON	

Struktur und Inhalt der Dimensionstabellen

Table_name	D_BLOCKING		D_SUPPL_TYPE		D_ORDER_TYPE	
Column_name	block_cd	block_desc	suppl_cd	suppl_desc	order_cd	order_desc
Inhalt	1	gesperrt	A	Post	100	Standard
	2	nicht gesperrt	B	DHL	200	Automatisch
	3	anteilig gesperrt	C	Paketshop	300	Mail
			D	UPS	400	Webstore
			E	Spedition	500	Telefon
			F	Selbstabholer	600	Fax

Table_name	D_COUNTRY		D_REGION		D_SEASON	
Column_name	country_cd	country_desc	region_cd	region_desc	season_cd	season_desc
Inhalt	DEU	Deutschland	S	Süd	4711	Frühling
	ITA	Italien	N	Nord	4712	Ostern
	FRA	Frankreich	O	Ost	4713	Weihnachten
	NED	Holland	W	West		
	DAN	Dänemark	NW	Nordwest		
			NO	Nordost		
			SW	Südwest		

Struktur und Inhalt der Faktentabelle, die leere Attribute enthalten.

F_SALES								
sales_id	article_id	sales_date	s_trans_date	book_date	sales_qty	block_cd	order_cd	season_cd
1000	12345678	01.10.2014	15.10.2014	05.10.2014	100	1	100	4711
1001	13580246	02.10.2014	15.10.2014		200	2		4713
1002	14938270	02.10.2014	15.10.2014	05.10.2014	300	2	500	4712
1003	16432097	02.10.2014	15.10.2014		400	3	200	4712
1004	18075307	03.10.2014	15.10.2014	05.10.2014	200	2		4712
1005	19882838	04.10.2014	15.10.2014		1000			4711
1006	21871122	05.10.2014	15.10.2014	05.10.2014	5	2	600	
1007	24058234	05.10.2014	15.10.2014	05.10.2014	10	1	300	4713

F_ORDERS							
order_id	article_id	order_date	o_trans_date	book_date	order_qty	country_cd	region_cd
11111	12345678	01.10.2014	15.10.2014	05.10.2014	500	DEU	S
22222	13580246	02.10.2014	15.10.2014		700	DEU	N
33333	14938270	02.10.2014	15.10.2014	05.10.2014	350	DEU	SW
44444	16432097	02.10.2014	15.10.2014		800	FRA	
55555	18075307	03.10.2014	15.10.2014	05.10.2014	1000	ITA	
66666	19882838	04.10.2014	15.10.2014		5000	ITA	O
77777	21871122	05.10.2014	15.10.2014	05.10.2014	20	DAN	
88888	24058234	05.10.2014	15.10.2014	05.10.2014	100		S

F_RECEIPTS						
recpt_id	article_id	recpt_date	r_trans_date	book_date	recpt_qty	suppl_cd
10	12345678	10.10.2014	15.10.2014	05.10.2014	15	A
20	13580246	10.10.2014	15.10.2014		200	
30	14938270	05.10.2014	15.10.2014	05.10.2014	100	A
40	16432097	05.10.2014	15.10.2014		570	C
50	18075307	05.10.2014	15.10.2014	05.10.2014	750	C
60	19882838	10.10.2014	15.10.2014		3300	B
70	21871122	10.10.2014	15.10.2014	05.10.2014	10	B
80	24058234	01.10.2014	15.10.2014	05.10.2014	90	

Die Dimensionstabellen sind bereits im DWH vorhanden. Eine neue Lieferung der Fakten wird z.B. am 15.10. (transfer_date) ins DWH geladen.

In allen drei Lieferungen fehlen einige Informationen. Beispielsweise ist es nicht eindeutig, mit welcher Bestellart (order_type) der Datensatz 1001 in der „Verkaufstabelle“ erfasst wurde.

Wird dann ein Report generiert, der die Mengenwerte der Faktentabelle, article_id und die Beschreibung der einzelnen Attribute im Monat Oktober (201410) beinhalten soll, dann würden die folgenden Records aus der Faktentabelle in dem Report fehlen.

Tabellenname	**Gesamt menge**	**Fehlende Records**				**Summe der fehlenden Mengen**
F_SALES	2215	1001	1004	1005	1006	1405
F_ORDERS	8470	4444	5555	7777	8888	1920
F_RECEIPTS	5035	20	80			290

Ein ähnliches SQL für die Sales-Tabelle**,** wie unten**,** würde dann im Hintergrund generiert und ausgeführt.

```
SELECT f.article_id
      ,d1.block_desc
      ,d2.season_desc
      ,d3.order_desc
      ,SUM(f.sales_qty) as SUM_SALES_QTY
  FROM F_SALES f
 INNER JOIN D_BLOCKING d1
    ON f.block_cd = d1.block_cd
 INNER JOIN D_SEASON d2
    ON f.season_cd = d2.season_cd
 INNER JOIN D_ORDER_TYPE d3
    ON f.order_cd = d3.order_cd
 WHERE 1=1
 AND to_char(f.sales_date ,'YYYY-MM') ='2014-10'
 GROUP BY f.article_id
         ,d1.block_desc
         ,d2.season_desc
         ,d3.order_desc;
```

Das Ergebnis des Reports könnte wie folgt aussehen. Es fehlen die Datensätze, bei denen die Werte für die Attribute „block_cd“, „order_cd“ und „season_cd“ fehlen.

Report				
article_id	block_desc	order_desc	season_desc	sum_sales_qty
12345678	gesperrt	Standard	Frühling	100
14938270	nicht gesperrt	Telefon	Ostern	300
16432097	anteilig gesperrt	Automatisch	Ostern	400
24058234	gesperrt	Standard	Weihnachten	10
			Summe	**810**

Zur Verbesserung der Datenqualität sollten einige Routinen erstellt werden, die als ein Teil im ETL-Prozess z.B. in der „Cleansing area“ auszuführen sind. Dieser „Self-Cleaning“-Mechanismus sollte in der Lage sein, die notwendigen Informationen aus dem Data-Dictionary der Datenbank zu sammeln und anhand bestimmter Regeln ein dynamisches SQL zu erstellen, nach deren Ausführung die fehlenden Attribute durch die zuvor definierten Defaultwerte eingefügt oder ersetzt werden.

Dabei wird einmalig eine Parametertabelle benötigt, die die Datengruppe, die Tabellen- und Spaltennamen, das Format der Spalten und die Defaultwerte beinhalten soll.

P_PARAM_TAB							
group_id	**table_name**	**colum_name**	**column_dec_val**	**column_char_val**	**column_date_val**	**default_desc**	**dml_type**
MASTERDATA	D_BLOCKING	BLOCK_CD	-1			UNKNOWN	I
MASTERDATA	D_COUNTRY	COUNTRY_CD		@		UNKNOWN	I
MASTERDATA	D_ORDER_TYPE	ORDER_CD	-1			UNKNOWN	I
MASTERDATA	D_REGION	REGION_CD		@		UNKNOWN	I
MASTERDATA	D_SEASON	SEASON_CD	-1			UNKNOWN	I
MASTERDATA	D_SUPPL_TYPE	SUPPL_CD		@		UNKNOWN	I
ORDERS	F_ORDERS	COUNTRY_CD		@		UNKNOWN	U
ORDERS	F_ORDERS	REGION_CD		@		UNKNOWN	U
RECEIPTS	F_RECEIPTS	SUPPL_CD	-1			UNKNOWN	U
SALES	F_SALES	BLOCK_CD	-1			UNKNOWN	U
SALES	F_SALES	ORDER_CD	-1			UNKNOWN	U
SALES	F_SALES	SEASON_CD	-1			UNKNOWN	U
SALES	F_SALES	BOOK_DATE			01.01.2999	UNKNOWN	U

Sie könnte manuell oder mit Hilfe eines Migrationsskripts durch die vorhandenen Informationen im Data-Dictionary und im DWH wie z.B. aus den Schedule- oder den Archivierungstabellen gefüllt werden. Die Dimensionstabellen sollen „insertable“ und die Faktentabellen „updateable“ sein.

Die Defaultwerte für die numerischen und alphanumerischen Spalten sowie für die Datumsfelder und deren Beschreibungen (default_desc) können beliebig festgelegt werden.

Die folgenden Routinen werden benötigt.

1. insert_default_value

Die erste Routine wird dazu benötigt, einen Defaultwert in der entsprechenden Referenz- bzw. Dimensionstabelle einzufügen, wenn dort keine Defaultwerte vorhanden sind. Dabei wird in einem Cursor anhand der Informationen aus dem Data-Dictionary und der Parametertabelle unabhängig von den Tabellen- und Spaltennamen die notwendigen Informationen gesammelt. Das erstellte dynamische SQL kann dann sofort per (EXECUTE IMMEDIATE) ausgeführt werden.

Normalerweise sollen die Referenztabellen nur ID und Description enthalten. Es gibt aber Datenmodelle, bei denen dies nicht der Fall ist. Deswegen muss man darauf achten, nur diejenige Dimensionstabellen in Betracht zu ziehen, die keine weiteren „NOT NULL“ Spalten außer ID und Beschreibung besitzen. Sonst kann das dynamische SQL nicht feststellen, für welche Tabellen welche „Not Null“-Spalten mit welchem Werte zu ersetzen sind.

```
FOR rec1 IN cursor_1 LOOP
  IF rec1.column_dec_value IS NOT NULL
    THEN
       lv_sql :=           ' INSERT INTO ' || rec1.table_name;
       lv_sql := lv_sql || ' SELECT ' || rec1.column_dec_value || ','''
                                      || rec1.default_desc ||'''';
       lv_sql := lv_sql || ' FROM ' || 'p_param_tab';
       lv_sql := lv_sql || ' WHERE NOT EXISTS (SELECT 1 FROM '
                                       || rec1.table_name || ' AS S2';
       lv_sql := lv_sql || '   WHERE ' || rec1.column_name || '='
                                       || rec1.column_dec_value || ';';
       . . . . . .
```

2. update_default_value

Die zweite Routine sollte dafür sorgen, alle leeren Spalten der aktuell gelieferten Stammdatentabellen (S_SUPPLIER, S_ARTICLE, S_ADVERTISING) und/oder Faktentabellen (F_RECEIPTS, F_SALES und F_ORDERS), die in der Parametertabelle aufgenommen worden sind, mit Defaultwerten zu belegen. Dabei werden nur die „U“-Sätze aus der Parametertabelle selektiert. Je nach dem welches Format die Spalte besitzt, wird dann das Update-Statement zusammengestellt.

```
IF rec1.column_dec_value IS NOT NULL
    THEN
       lv_sql :=           ' UPDATE ' || rec1.table_name;
       lv_sql := lv_sql || '    SET ' || rec1.column_name
                        || '= ' || rec1.column_dec_value ;
       lv_sql := lv_sql || '  WHERE ' || rec1.column_name || ' IS NULL; ';

  . . . . . .
```

Das erstellte Update-Statement kann dann sofort per (EXECUTE IMMEDIATE) ausgeführt werden. Die Routine kann für bestimmte Datengruppe aktiviert oder deaktiviert werden.

3. insert_new_attributes

Eine weitere Routine zur Wartung der Parametertabelle sollte ebenfalls realisiert werden, die dafür sorgt, alle neuen Faktentabellen und die entsprechenden Attribute, die z.B. nach einem neuen Release eingeführt worden sind, anhand bestimmter Regeln automatisch in der Parametertabelle hinzuzufügen. Dies erleichtert eine manuelle Pflege der Stammdaten- oder Faktentabellen in der Parametertabelle.

Zunächst werden in einem Cursor nur die „I“-Datensätze („Insertable“) aus der Parameter-Tabelle gesucht, für die dort kein „U“-Satz vorhanden ist. Dann sollen die neuen Informationen der Faktentabelle als „U“-Satz in die Parameter-Tabelle eingefügt werden. Die bisherigen Regeln bleiben unverändert.

4. MigrationScript

Um eine manuelle Pflege der Parametertabelle zu sparen, ist es empfehlenswert, ein Migrationsskript zu erstellen, um die betroffenen Tabellen und Spalten in die Parametertabelle aufzunehmen.
Die Aufnahme der betroffenen Spalten und Tabellen in der Parametertabelle kann Schritt für Schritt geschehen, in dem man sie als Parameter an diese Routine übergibt.

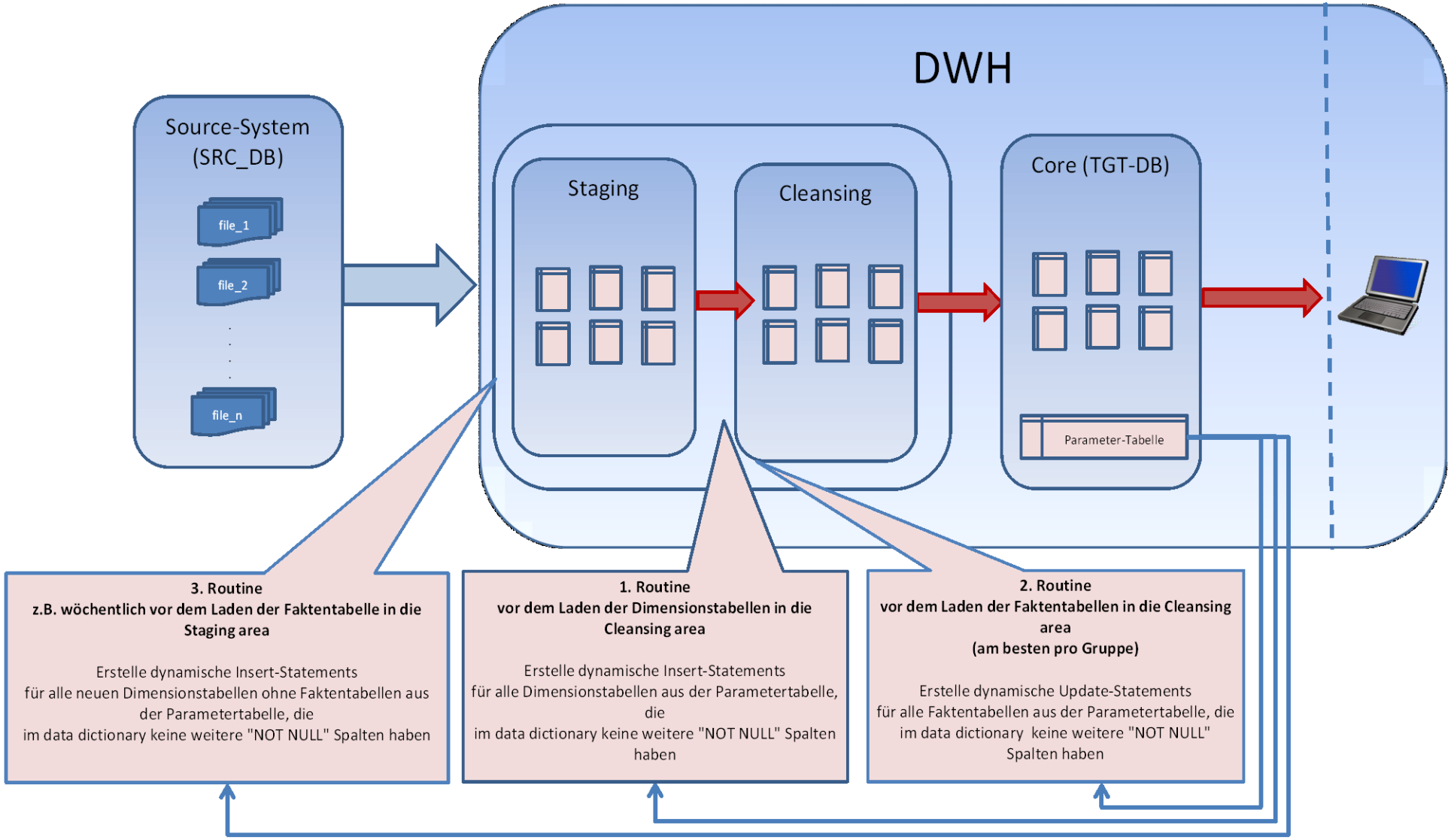
Bemerkungen

Da es bei den ETL-Prozessen üblich ist, die zusammengehörigen Faktentabellen in einer eigenen Gruppe zu laden (z.B. alle Informationen zu Abverkäufe werden in der Gruppe „SALES“ geladen) , ist es empfehlenswert, den Namen dieser Gruppe an die neuen Routinen zu übergeben. Dadurch kann man das Update der aktuell gelieferten Faktentabellen parallel durchführen.

Das Ergebnis des Reports könnte wie folgt aussehen, wenn alle leer gelieferten Attribute durch diese Methode mit den entsprechenden Defaultwerte gefüllt werden.

Report				
article_id	block_desc	order_desc	season_desc	sum_sales_qty
12345678	gesperrt	Standard	Frühling	100
13580246	UNKNOWN	UNKNOWN	UNKNOWN	200
14938270	nicht gesperrt	Telefon	Ostern	300
16432097	anteilig gesperrt	Automatisch	Ostern	400
18075307	UNKNOWN	UNKNOWN	UNKNOWN	200
19882838	UNKNOWN	UNKNOWN	UNKNOWN	1000
21871122	UNKNOWN	UNKNOWN	UNKNOWN	5
24058234	gesperrt	Standard	Weihnachten	10
			Summe	**2215**

Das folgende Bild zeigt schematisch, an welcher Stelle des ETL-Prozesses die neuen Routinen zum Einsatz kommen können.



Fazit:

- Mit dieser Methode ist man sehr flexibel, die leer gelieferten Attribute mit Defaultwerten zu belegen. Bei einer Änderung in der Parametertabelle könnte man das Verhalten zum Default-Handling leicht beeinflussen. Diese Funktionalität könnte jeder Zeit für bestimmte Gruppen der Stammdaten- oder Faktentabellen durch einen Schalter ein- und ausgeschaltet werden, falls aus technischen oder fachlichen Gründen nicht erwünscht ist, die Attribute bei allen Tabellen zu ändern.

- Man könnte meinen, beim Einfügen einer NVL-Funktion in jeder View dieses Problem umgehen zu können.

  Dazu habe ich die folgenden Gegenargumente:

  a. Man muss bei allen betroffenen Views die entsprechenden Attribute um die NVL-Funktion erweitern. Es stellt sich die Frage, ob erstens bei der Umstellung eine oder mehrere Views nicht doch übersehen wurden und zweitens kein Eingabefehler bei der Pflege der Views gemacht wurde.

  b. Die hardcodierten Werte für die Defaultwerte sind nicht immer beliebt.
  NVL(d1.attr_tab_1_id , -1)
  NVL(d1.block_desc , 'Unknown')

  c. Falls aus irgendeinem Grund die Defaultwerte geändert werden sollten, dann müssen alle Views angepasst werden, was zu Mehraufwand führt.

- Diese Methode kann für alle Stammdaten- und/oder Faktentabellen eingesetzt werden, bei denen die entsprechenden Attribute mit NULL-Werte an das DWH geliefert werden.

  Sollten die Attribute in den Source-Systemen andere Defaultwerte als „NULL“ zugewiesen bekommen, wie z.B. ‚@@@‘, dann funktioniert diese Methode nicht. In diesem Fall sollten durch die entsprechenden Änderungen die Routinen noch flexibler werden, was zu Mehraufwand führt.

- Bei dieser Methode handelt es sich um Erfahrungen aus langjährigen Kundenprojekten in der Handelsbranche. Dieser Mechanismus funktionierte im Delta-Load mit großen Datenmengen ohne bemerkbare Performance-Einbußen. Da aber für ein Initial-Load keine Erfahrungswerte vorliegen, sollte hierzu eine andere Methode in Betracht gezogen werden.

Kontaktadresse:
Feraydoun Mohajeri
Triestram & Partner GmbH
Kohlenstraße 55
D-44795 Bochum

f.mohajeri@t-p.com

Fon: +49 (0)234 943 750
Fax: +49 (0)234 452 206
Internet: http://www.t-p.com/