Fehlertoleranz und Robustheit von ETL-Prozessen

Christian Borghardt
areto consulting gmbh
Köln

Schlüsselworte:

Datenbank, Error-Tabellen, Historisierung, Datenqualität

Einleitung

In einem Datawarehouse werden Daten häufig zum ersten Mal konsolidiert gespeichert und können mit Stammdaten aus anderen Quellen verglichen werden. Dies hat zur Folge, dass die Qualität von Daten häufig erst hier bekannt wird und diese meist unter den Erwartungen bleibt. Um nachhaltig die Datenqualität zu steigern muss dazu Feedback an die Quellen geliefert werden, welche die Daten angleichen, sodass die Stammdaten nicht nur im Datawarehouse konsolidiert sind, sondern auch eine Konsolidierung in der Quelle stattfinden.

Ebenfalls können in den Quellsystemen Eingaben getätigt werden, welche im Datawarehouse zu Fehlern führen. Beispielhaft können Spalten fehlerhaft werden oder Felder, welche im Warehouse gefüllt sein müssen, also einen NOT NULL Constraint besitzen, keinen Eintrag erhalten.

Error-Tabellen

Damit die Prozesse zur Beladung des Datawarehouses nicht bei jedem Fehler direkt abstürzen, existiert in einer Oracle Datenbank ab Version 10g Release 2 die Möglichkeit des Error-Loggings. Dazu muss zunächst eine Error-Tabelle auf der Datenbank erzeugt werden. Dazu existiert ein Package, welches von Oracle bereitgestellt wird. Dies erzeugt die Error-Tabelle und befüllt diese bei einem Fehler nach folgendem Muster:

- ORA_ERR_NUMBER$, hier wird die Fehlernummer eingefügt
- ORA_ERR_MESG$, hier wird der Fehlertext eingefügt
- ORA_ERR_ROWID$, falls es sich um ein Update handelt, wird die ROWID eingetragen
- ORA_ERR_OPTYP$, hier wird die Art der DML eingetragen
- ORA_ERR_TAG$, hier kann ein Freitext eingetragen werden
- Spalten aus Tabelle ohne Constraints und alle haben den Typ Varchar2(4000)

Die Existenz der Error-Tabelle genügt jedoch noch nicht um potentielle Fehler abzufangen. In unseren Statements oder automatischen Prozessen durch ETL-Tools muss eingestellt werden, dass eine bestimmte Anzahl an Fehlern in der Error-Tabelle landen kann bis dieser abbricht. Für Statements muss folgendes hinzugefügt werden:

```
<STATEMENT> …
LOG ERRORS INTO <Name der Error-Tabelle> ('<ORA_ERR_TAG>')
REJECT LIMIT UNLIMITED / <X>;
```

Historisierung und Wiederanlauffähigkeit

Im zweiten Teile gehe ich auf die Wiederanlauffähigkeit ein, die eine nachträgliche Ladung und/oder Korrektur zulässt. Dazu werden zwei Historisierungsverfahren präsentiert: Slowly Changing Dimensions Typ 2 (SCD2) und die Snapshot-Historisierung.

In einer SCD2 historisierten Tabelle existiert zu einem Abfragezeitpunkt immer nur maximal ein gültiger Datensatz. Um dies zu gewährleichten, gibt es zwei Spalten, die den Beginn (<Gueltig_Ab>) und das Ende (<Gueltig_Bis>) der Gültigkeit angeben. Wenn eine Abfrage auf die Tabelle erfolgt, dann muss die Einschränkung gemacht werden:

```
<DATUM> BETWEEN <Gueltig_Ab> and <Gueltig_Bis>
```

Fehlerhafte Datensätze können dann bei einer neuen Ladung abgeschlossen werden und der neue Datensatz kann eingefügt werden.

Für die Snapshot-Historisierung kann pro fachlichem Ladedatum (<Ladedatum>) eine Partition angelegt werden und es muss bei Abfragen die Einschränkung gemacht werden:

```
<DATUM> = <Ladedatum>
```

Falls hier eine fehlerhafte Anlieferung statt findet, wird die alte Partition gedroppt und die korrigierten Daten können so eingespielt werden.

Fazit

Durch die Error-Tabelle kann man technische Fehler abfangen und die Fehlersuche wir stark vereinfacht, da die fehlerhaften Daten bereits gefunden sind und nicht erst in der Analyse gesucht werden müssen. Durch die Fehlerhistorie können wir ebenfalls sehen, ob wir systematische Fehler erkennen und gegebenenfalls Prozesse anpassen müssen oder die Quelle korrigieren sollten.

Fehler sollten so früh wie möglich erkannt werden, da eine nachträgliche Ladung immer mehr Aufwand bedeutet und bei Surrogat Integer Keys (SIC, künstliche Schlüssel) müssen die Referenzen ebenfalls nachträglich geladen werden.

Ausblick

Für die SCD2 gibt es noch die Möglichkeit von Lücken-oder Embryodatensätzen, welche die Stabilität erhöhen und die Nachladbarkeit vergangener Tage ermöglichen.

Es existieren noch andere technische Fehler, welche auch die Prozesse stören können, wie z.B. eine Abfrage über den Tablespace oder andere Datenbank Parameter. Dies kann in einem Dashboard realisiert werden, welches zu jedem Zeitpunkt diese Informationen ausliest und so eine Möglichkeit liefert auf einem Blick die Beladungen zu beobachten.

Christian Borghardt

areto consulting gmbh
Julius-Bau-Str. 2
D-51063 Köln

Telefon: +49 (0) 221 66 95 75-0
Fax: +49 (0) 221 66 95 75-99
E-Mail Christian.Borghardt@areto-consulting.de
Internet: www.areto-consulting.de