Datenqualität mit Oracle-Produkten

Götz Gleitsmann
ORBIT Gesellschaft für Applikations- und Informationssysteme mbH
Bonn

Schlüsselworte
Datenqualität, Generische Mappings, Oracle Data Integrator, Oracle Warehouse Builder, Informatica PowerCenter, Oracle Application Express, Einbindung von Drittlösungen

Einleitung

Gute Datenqualität (DQ) ist eine notwendige Bedingung für den reibungslosen Ablauf von IT-Prozessen. Schlechte Datenqualität kann hingegen zu massiven Fehlern führen und großen wirtschaftlichen Schaden anrichten. Jedes Unternehmen, das mit Daten zu tun hat – und welches hat das nicht? – muss sich Gedanken über die Qualität dieser Daten machen. Die wenigsten tun das.
Inhaltliche Fehler können durch Abweichung von Adressformaten, Abkürzungen, Rechtschreibfehler, veraltete Information oder vertauschte Vor- und Nachnamen entstehen. Allgegenwärtig ist auch die Gefahr der Entstehung von Dubletten. Hierzu reicht bereits ein falscher Buchstabe oder ein führendes Leerzeichen aus. Daher muss festgestellt werden, ob unterschiedliche Personen, Firmen oder andere Entitäten nicht in Wirklichkeit gleich sind
In dieser Präsentation werden Möglichkeiten aufgezeigt, wie fehlerhafte Datensätze in der ETL-Strecke durch die Verwendung vordefinierter SQL-Statements in generischen Mappings identifiziert werden können. Auch werden die intrinsischen DQ-Möglichkeiten der Oracle-Produkte Data Integrator (ODI), Warehouse Builder (OWB), Application Express (APEX) sowie von Informatica vorgestellt. Auf technische Details wird hier bewusst verzichtet. Damit kann ein Weg aufgezeigt werden, wie mit Hilfe dieser Produkte die DQ-Anforderungen des Kunden umgesetzt werden können.

Generische Mappings

Sind die DQ-Anforderungen vielseitig, kommen eigentlich nur generische Mappings in Betracht, die in einem ETL/ELT-Tools angesiedelt sind und aus den zu prüfenden Tabellen alle als fehlerhaft erkannten Datensätze auslesen. Sie enthalten die Prüflogik in Form eines SQL-Skripts und schreiben die wichtigsten Felder einheitlich in Sondertabellen. Bei einem Siebel CRM-System wären dies die Felder ROW_ID, CREATED, CREATED_BY, LAST_CHANGED, LAST_CHANGED_BY, TABLE_NAME, ENTITY, CHECKING_DATE, ETL_RUN_NUMBER, DQ_ERROR_NUMBER, DQ_ERROR_TYPE, ERR_COL_NAME und ERR_COL_CONTENT. Hinzu kommen je nach Bedarf weitere Felder. Die Sondertabellen enthalten damit alle wichtigen Informationen, die für die Korrektur der fehlerhaften Daten notwendig sind. Hier ist ein Code-Beispiel gezeigt, in dem in einer Siebel CRM-Tabelle das Vorhandensein der Anrede „Herr“ bzw. „Frau“ geprüft wird.

```
SELECT
      s.row_id, s.created, s.created_by, s.last_upd, s.last_upd_by,
      'S_CONTACT s, CX_PARTY, S_ORG_EXT, S_ORG_EXT_X, CX_DIVISION',
      'Ansprechpartner', 1, S_CONTACT.fst_name || ' ' ||
      S_CONTACT.last_name, x_address_salut, x_letter_salut, attrib_46,
      sysdate, 0, 999, 'Sonderregel Anrede', 'X_ADDRESS_SALUT',
     X_ADDRESS_SALUT
FROM
      S_CONTACT, CX_PARTY, S_ORG_EXT, S_ORG_EXT_X, CX_DIVISION
WHERE
     S_CONTACT.PR_DEPT_OU_ID = S_ORG_EXT.ROW_ID AND S_ORG_EXT.ROW_ID =
     CX_PARTY.ACCOUNT_ID AND CX_DIVISION.ROW_ID =
     CX_PARTY.ORGANIZATION_ID AND S_ORG_EXT.ROW_ID =
     S_ORG_EXT_X.PAR_ROW_ID (+) AND CX_PARTY.TYPE = 'Account Unit' AND
     emp_flg = 'N'
AND
      (
      instr(lower(x_address_salut), 'herr ') = 0 and
      instr(lower(x_address_salut), 'frau ') = 0 or
      instr(lower(x_letter_salut), 'herr ') = 0 and
      instr(lower(x_letter_salut), 'frau ') = 0
      )
```

In Informatica PowerStation sind generische Mappings problemlos möglich. Man kann sogar das gleiche generische Mapping mehrmals nacheinander aufzurufen und ihm immer andere Parameterdateien mitzugeben, in denen das für die DQ-Prüfung relevante SQL enthalten ist. So kann die ETL-Strecke flexibel an sich ändernde DQ-Anforderungen angepasst werden, ohne die Vorgaben der Best Practice formal zu verletzen.
In OWB sind generische Mappings ab der Version 11gR2 möglich, denn es kann auch hier ein SQL Override definiert werden. Hingegen bietet ODI die Möglichkeit generischer Mappings nicht. Mit der Option „ODI Procedures“ ist dennoch eine flexible DQ-Modellierung möglich, denn das Prüf-SQL kann in Form einer gespeicherten Prozedur verpackt und mit ODI Procedures aufgerufen werden.

Zusatzoptionen: Oracle Data Integrator

Neben der Definition generischer Mappings enthalten die ETL/ELT-Tools auch intrinsische Möglichkeiten, die nur einen Teil des gesamten DQ-Spektrums abdecken, im Einzelfall aber eine sehr sinnvolle Hilfe sein können. So bietet ODI mehrere Möglichkeiten. Zunächst können als fehlerhaft erkannte Daten der Quellsysteme vom ELT-Lauf ausgeschlossen werden; die betroffenen Datensätze werden dafür in eine Sondertabelle geschrieben.
Die sog. „Data Quality Firewall“ bietet einen ganzheitlichen Ansatz, der aus folgenden Schritten besteht:

1. Valide Daten durchlaufen die ELT-Strecke ohne Eingriff.
2. Invalide Datensätze werden in gesonderte Tabellen umgeleitet.
3. Zurückgewiesene Daten können anschließend überprüft werden.
4. Reparierte Daten können nachträglich weitergeleitet oder recycled werden.

Die für die Datenqualität benutzten Regeln werden im Metadaten Repository gespeichert. Sie können entweder durch Reverse Engineering aus der Datenbank ausgelesen werden (z.B. Constraints oder Fremdschlüssel) oder aber als frei definierte Regeln im Mapping Designer erstellt werden.

Mit Hilfe von ODI-Mappings lassen sich zahlreiche DQ-Regeln realisieren. Diese teilen sich auf in solche der Dublettenerkennung, der referenziellen Integrität und der Wertevalidierung. Die Dublettenerkennung setzt auf bestimmten Merkmalen auf wie etwa gleichen E-Mail-Adressen oder gleicher Produktbezeichnung bei verschiedenen Bestellnummern. Entsprechende Regeln werden in ODI in Form von Primär- oder Ersatzschlüsseln bzw. Unique Constraints implementiert. Die Regeln der referenziellen Integrität beschäftigen sich hingegen mit Kunden ohne zugeordneten Account Manager oder Bestellungen invalider Kunden. Hier greifen Schlüsselvergleiche, also WHERE-Bedingungen bzw. komplexere Funktionen wie Spalte A = f(Spalte B, Spalte C). Bei der Wertevalidierung schließlich wird die Konsistenz auf Datensatzebene überprüft. Als Beispiele seien fehlende Postleitzahlen oder Bestellungen mit formal falscher E-Mail-Adresse genannt. Auch die Wertevalidierung wird über SQL-Bedingungen implementiert.

Mit ODI kann die Datenqualität an folgenden Stellen überwacht werden:

1. Innerhalb der Quellsysteme;
2. auf der ELT-Strecke;
3. Innerhalb des Data Warehouse.

In allen drei Fällen werden ODI-Mappings verwendet, in denen die entsprechende Logik enthalten ist. Innerhalb der Quellsysteme bzw. des DWH sorgen automatisierbare Prozesse dafür, dass in regelmäßigen zeitlichen Abständen fehlerhafte Daten aus den regulären Tabellen in Sondertabellen geschrieben werden. Wird auf der ELT-Strecke die Datenqualität überwacht, so geschieht dies, indem fehlerhafte Datensätze nicht in das DWH geschrieben sondern ebenfalls in spezielle Tabellen umgeleitet werden. Als vierte Möglichkeit, zusätzlich zu den o.g. drei anderen, können als fehlerhaft erkannte Datensätze aus dem DWH wieder in die ELT-Strecke eingebracht werden.
Die Fehlertabellen können entweder innerhalb der ODI-Umgebung (Designer-Fenster) betrachtet werden oder aber mit jedem beliebigen Datenbanktool, z.B. SQL Developer.

Zusatzoptionen: Oracle Warehouse Builder

Die Stärken des OWB liegen in der Bereinigung von Adressen und der Entfernung von Dubletten. Für jeden dieser beiden Schritte ist ein Transformationsoperator im Funktionsumfang enthalten. Beide Operatoren haben definierte Eigenschaften und können in Mappings eingebaut werden.
Der Operator „Name and Address“ korrigiert inhaltliche Fehler durch

1. Parsen oder Trennen von Namens- und Adressdaten in einzelne Elemente;
2. Postalisch korrektes Standardisieren von Abkürzungen, Titeln und Adressen;
3. Validieren und Korrigieren von Adress- und Ortsinformationen;
4. Anreicherung mit Zusatzinformationen (Geschlecht, PLZ, Landesvorwahl etc.).

In diesem Operator werden die Adressregeln modelliert. Der zugrunde liegende „Name and Address Server“ wird bei Installieren der Oracle-Datenbank mit installiert.
Für die eigentliche Benutzung der Datenqualitätsoptionen in OWB ist die Einbindung von Fremdsoftware und Adressbibliotheken erforderlich, die für die Verwendung mit OWB zertifiziert sein müssen. Sie brauchen sehr viel Speicherplatz, der in der Größenordnung mehrerer GB liegen kann. Als Beispiele seien Trillium oder First Logic genannt. Eine Übersicht zu allen Anbietern lässt sich über folgende Seite abrufen:

www.oracle.com/technology/products/warehouse/htdocs/OTN_Partners.html

Die Adressbibliotheken enthalten aktuelle in- und ausländische Adressen sowie korrekte Schreibweisen und Abkürzungen. Mit ihrer Hilfe werden fehlende Angaben ergänzt und veraltete oder falsche Daten (z.B. fehlerhafte Zuordnung einer Postleitzahl) korrigiert. Dies geschieht zur Laufzeit unter Beachtung der im Operator „Name and Address“ modellierten Regeln.

Beispiel: der Datensatz „Fritz Schmitz|Fliegenbusch 4711|Krefeld-Hüls“ wird automatisch geändert in „Fritz|Schmitz|Am Fliegenbusch|4711|Krefeld|Deutschland|47839“.

Man sieht, dass Vor- und Nachname voneinander getrennt und automatisch eine Postleitzahl ergänzt wurde. Auch wurde Krefeld-Hüls in Krefeld geändert, da der Stadtteil in der Postleitzahl enthalten ist.

Match Merge. Die zweite Säule des DQ-Pakets im Warehouse Builder ist der Operator „Match Merge“. Hiermit können sowohl in strukturierten als auch in unstrukturierten Daten Dubletten ausfindig gemacht und anschließend zusammengeführt werden. Die Adaptierung der Match- und Merge-Regeln erfolgt mit Hilfe eines intuitiv zu bedienenden Assistenten. Dabei stehen vordefinierte elementare Match-Regeln zur Verfügung, z.B.:

1. Strip Noise Words
2. Match on abbreviations
3. Match on acronyms
4. Detect switched name order
5. Match on initials
6. Detect compound name.

Um die Dublettenerkennung möglichst sicher zu machen, sollten die Daten zuvor bereits mit dem Operator „Name and Address“ bereinigt worden sein. Der Operator „Match Merge“ besteht aus einer Eingabe- und zwei Ausgabegruppen. Er ist architektonisch gesehen statisch, denn es können keine Gruppen hinzugefügt oder entfernt werden. Beide Ausgabegruppen, nämlich MERGE (mit den konsolidierten Daten) und optional XREF (mit der Dokumentation des Konsolidierungsprozesses) werden im PL/SQL-Format generiert. Letztere kann durch den Entwickler den Anforderungen angepasst werden.
Bei der Ausführung eines OWB-Mappings mit eingebautem Match-Merge-Operator ist allein der Betrieb im zeilenbasierten Modus möglich. Werden mehrere Regeln definiert, so erfolgt die Auswertung zur Laufzeit nach der Oder-Logik. Zur Steigerung der Performance lassen sich die Daten mit Hilfe sog. „Match bins“ gruppieren, bei denen nur begrenzte Bereiche durchsucht werden und nicht die ganze Tabelle. Beispielweise könnten Kundenadressen nach Stadtnamen oder Postleitzahlen gruppiert werden. Insbesondere sind solche Match Bins dann sinnvoll, wenn aus den o.g. elementaren Regeln komplexe benutzerdefinierte und damit „langsame“ Regeln erstellt worden sind.
Sobald die Dubletten heraus gefunden sind, kann mit Hilfe von insgesamt elf Merge-Regeln eine Konsolidierung erfolgen. Soll etwa der längste Vorname als konsolidierter Datensatz „überleben“, so wird die Regel „Min Max Record“ verwendet. Anderes Beispiel: mit TAKE_VERIFIED_ADDRESS wird festgelegt, dass der Datensatz verwendet werden soll, der auch in der Postdatenbank (Drittanbieter) zu finden ist.

Informatica Analyst und Developer
Informatica bietet die Möglichkeit des Data Profiling. Mit dem Analysten können Tabellen auf ihr Wertespektrum untersucht und die Häufigkeit bestimmter Ausprägungen ermittelt werden. Auch ist es möglich, Muster zu erkennen. Zum Beispiel kann man Spalten, die Datumswerte im String-Format speichern, darauf untersuchen, ob und wie viele Datensätze es gibt, deren Format nicht „dd.mm.yyyy“ sondern abweichend aufgebaut ist.
Der Developer bietet erweiterte Möglichkeiten. Insbesondere lassen sich dort Mappings entwickeln, mit denen Tabellen auf fehlerhafte Datensätze untersucht werden. Diese Mappings können mehrere Quelltabellen auslesen; bei der Gestaltung der Join- und Filterbedingung ist man völlig frei und kann damit sämtliche DQ-Anforderungen abdecken. Dabei sind die Gestaltungsmöglichkeiten mit denen aus Informatica PowerCenter (IPC) kompatibel. Bezüglich statistischer und DQ-spezifischer Operatoren gehen sie sogar deutlich über IPC hinaus.

Mit dem Informatica Developer erstellte Mappings kann man in dieser Umgebung für Ad-hoc-Ausgaben laufen lassen. Die als fehlerhaft erkannten Datensätze werden ausgegeben. Ist man später an einer fest installierten DQ-Strecke interessiert, können diese Mappings nach IPC exportiert werden. Dies unterliegt jedoch der o.g. Beschränkung auf das gemeinsame Spektrum der Operatoren.

Oracle Application Express
Bei der Erstellung von APEX-Anwendungen lassen sich für Eingabefelder Validierungen einstellen. Es können sowohl einzelne Felder als auch ganze Seiten validiert werden. Neben der „Not NULL"-Bedingung sind reguläre Ausdrücke oder String-Vergleiche möglich. Während reguläre Ausdrücke z.B. die korrekte Formatierung von Telefonnummern überwachen können, helfen String-Vergleiche, unzulässige Zeichen zu unterbinden. Beschränkungen des Wertebereichs bei Zahlen werden über SQL oder PL/SQL definiert, ebenso die Validierungen ganzer Seiten. Die Erstellung der SQL- oder PL/SQL-basierten Validierungen erfolgen mit Hilfe eines Assistenten. Dort werden die Bedingungen (z.B. Gehalt > 0) eingegeben; das SQL wird anschließend daraus generiert. Ebenso erfolgt im Assistenten die Eingabe von Inhalt und Ort der Fehlermeldung.

Fazit
Die Eingabe falscher Daten kann zurückgedrängt werden, wenn bereits in der Eingabemaske eine Überprüfung stattfindet. Dies ist bei den Oracle-Produkten APEX und Siebel CRM möglich. Verwendet der Kunde ein CRM-System, das keine durchgehende referenzielle Integrität gewährleistet, sollte diesbezüglich eine Überwachung geschaffen werden. Dies kann entweder in Form generischer Mappings oder mit Hilfe der oben beschriebenen Funktionalitäten von ODI erfolgen. Liegen die vermuteten DQ-Probleme eher im Bereich der Kontaktdaten, so wäre das DQ-Paket des OWB das Mittel der Wahl. Abschließend sei hervorgehoben, dass es bei der Vielzahl an Produkten kein Patentrezept gibt. Vielmehr ist im Einzelfall abzuwägen, mit welchen Mittel das geforderte DQ-Ziel am besten erreicht werden kann.

Kontaktadresse:
Götz Gleitsmann
ORBIT Gesellschaft für Applikations- und Informationssysteme mbH
Mildred-Scheel-Str. 1
D- 53175 Bonn

Telefon: +49 228 95693-673
Fax: +49 228 95693-99
E-Mail goetz.gleitsmann@orbit.de
Internet: www.orbit.de