Toolgestützte Analysen in Data-Quality-Projekten mit dem Oracle Warehouse Builder

Stefan Schrickel, OPITZ CONSULTING GmbH

Datenqualität und Datenmanagement tauchen häufig in Umfragen zu Schwerpunktthemen in IT-Projekten auf [1]. Der Artikel gibt einen Einblick in die Methoden und Techniken des Datenqualitätsmanagements und zeigt dabei die Möglichkeiten der technischen Unterstützung durch den Oracle Warehouse Builder (OWB) auf.

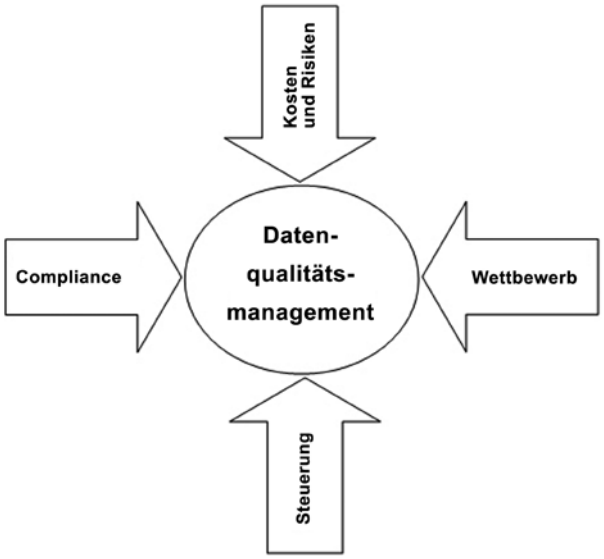

Abbildung 1: Spannungsfeld der Datenqualität

Der erste Teil liefert eine theoretische Einordnung des Themas „Datenqualität" und zeigt, wie schlechte Daten entstehen und an welchen Kriterien korrekte Daten festzumachen sind. Darüber hinaus wird der Begriff „Data Profiling" als grundlegende Methode in Data-Quality-Projekten erläutert, gefolgt von einem Überblick über die verschiedenen Analysemethoden. Ein Analysebeispiel stellt die Umsetzung mit dem OWB detailliert vor.

Wie entstehen „schlechte" Daten?

Mögliche Ursachen für schlechte Datenqualität sind im Spannungsfeld der Daten im Unternehmen zu finden (siehe Abbildung 1). Es lassen sich vier wesentliche Einflussfaktoren auf die Datenqualität feststellen [2].

In der Grafik ist bereits der Begriff „Datenqualitätsmanagement" zu sehen. Er verdeutlicht, dass die Verbesserung der Datenqualität in einem IT-Projekt oder im gesamten Unternehmen als Prozess anzusehen ist. Die Literatur [3] führt immer wieder folgende Kriterien für die Ansprüche an Datenqualität auf:

- *Verfügbarkeit*
  Die Daten müssen dem Anwender jederzeit und unverfälscht zur Verfügung gestellt werden
- *Vertrauenswürdigkeit*
  Die Verlässlichkeit der Datenquelle muss gegeben sein, damit die Daten als vertrauenswürdig angesehen werden können
- *Umfang*
  Daten müssen in ihrer Informationsbreite und -tiefe dem jeweiligen Anwendungszweck entsprechen
- *Übersichtlichkeit*
  Da wichtige Unternehmensentscheidungen aus der Interpretation der Daten abgeleitet werden, müssen die Informationen für den Anwender leicht verständlich und damit übersichtlich sein
- *Dokumentation*
  Nur eine umfangreiche Dokumentation der Daten macht es möglich, richtige Entscheidungen aus ihnen abzuleiten
- *Handhabbarkeit*
  Das Design von Softwareprodukten berücksichtigt auch ergonomische Gesichtspunkte. Dieser Punkt ist im Design eines Datenmodells zu berücksichtigen, denn nur eine schnelle und intuitive Navigation durch die Daten ermöglicht eine umfassende Auswertung
- *Aktualität*
  Der Informationsumfang eines Unternehmens wächst von Tag zu Tag und von Jahr zu Jahr. Die Betrachtung historischer Daten kann je nach Anwendungsfall von hoher Bedeutung sein. Als noch wichtiger gelten in den meisten Fällen die aktuellen Daten, die einem Unternehmen zur Verfügung stehen
- *Korrektheit*
  Dieser Aspekt wird landläufig am ehesten mit Datenqualität gleichgesetzt. „Zwei plus zwei" muss immer „vier" ergeben. Die Summe von Einzelposten ergibt in diesem Fall den richtigen Gesamtbetrag. Die Information muss also einfach „richtig" sein
- *Stimmigkeit*
  Die Stimmigkeit von Daten ergibt sich aus den Daten selbst und ihrem Kontext, beispielsweise sind die Daten einer Rechnung nur dann stimmig, wenn Rechnungsadresse, Positionen der Rechnung, Gesamtbetrag

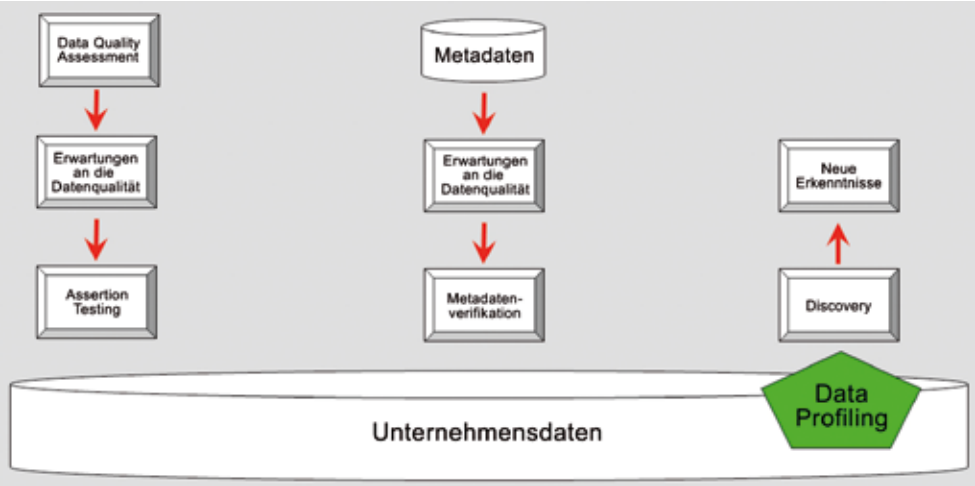

Abbildung 2: Vorgehensweisen für das Data-Profiling

etc. allesamt korrekt und vollständig vorhanden sind
- *Vollständigkeit*
  Im letzten Beispiel klang bereits der Begriff „Vollständigkeit" an. Für einen Anwendungsfall wie ein Controlling-System müssen daher immer auch Daten des Einkaufs, des Verkaufs und aller anderen Kosten oder Einnahmen eines Unternehmens vorliegen

Bei Beobachtungen zur Datenqualität in verschiedenen Projekten haben sich die folgenden beiden Ansätze zur Erkennung von Datenqualitätsmängeln herauskristallisiert:

- Top-Down-Methode: „Wir wissen und vermuten Dinge, die nicht stimmen"
- Bottom-Up-Methode: „Wir lassen uns überraschen, was da noch kommt"

Abbildung 2 zeigt, welche Methoden diesen beiden Ansätzen zugeordnet sind. Auf der linken Seite wird die Vorgehensweise der Top-Down-Methode dargestellt, zum einen mit Bezug auf die Daten selbst, zum anderen aus Sicht der Metadaten. Auf der rechten Seite findet sich mit dem Data Profiling eine mögliche Vorgehensweise der Bottom-Up-Methode. Auf diese Methode möchte ich im Weiteren näher eingehen.

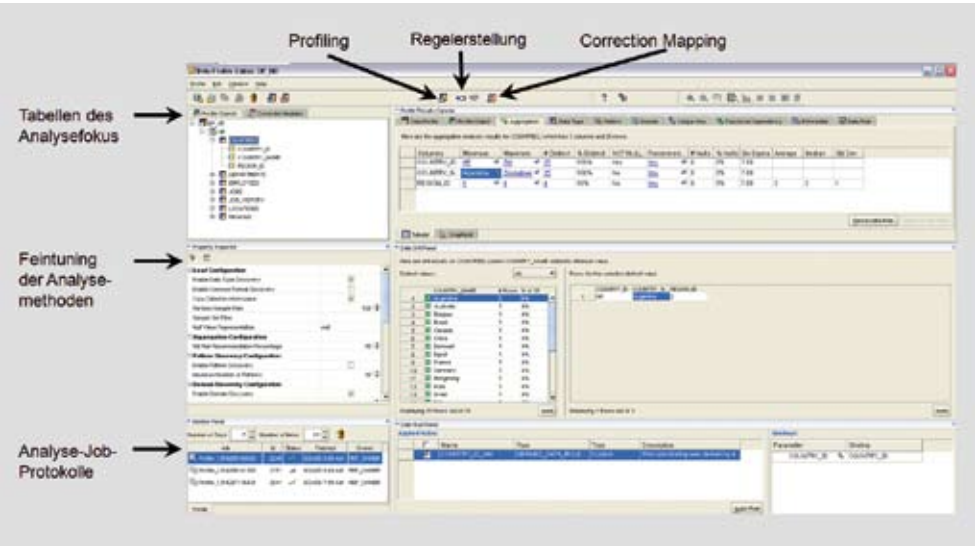

Abbildung 3: Data-Quality-Option des Oracle Warehouse Builder

Data Profiling – eine hilfreiche Methode zur Erkennung von Datenmängeln

Data Profiling stellt im Wesentlichen einen automatisierten Arbeitsschritt zur Untersuchung von Daten und deren Strukturen dar. Dieses Vorgehen hat sich in verschiedenen Praxisprojekten als guter Einstieg zur Erkennung und ersten Skizzierung von Datenqualitätsproblemen erwiesen.

Die Data-Quality-Option ist in den Oracle Warehouse Builder integriert und folgt daher im Design und in der Funktionalität der Gesamtanwendung. So können entsprechende Datenkorrekturmaßnahmen, die aus dem Profiling abgeleitet werden, direkt in einen ETL-Prozess eingebunden werden. Abbildung 3 zeigt die wichtigsten Bedienelemente.

Entscheidend für die einzelnen Analysen sind der Parameterdialog und der Reportbereich auf der rechten Seite. Für detaillierte Erläuterungen zum Tool und seiner Bedienung sind die Tutorials des Herstellers auf http://www.oracle.com/technology/obe/11gr2_db_prod/index/index.htm empfohlen.

Die Tabelle 1 zeigt, welche speziellen Analysemethoden der OWB bietet und wie diese in der Praxis eingesetzt werden können.

Die Ergebnismaske des Profilings wird nachfolgend am Beispiel der Musteranalyse vorgestellt. Die Erfahrung aus verschiedenen Datenqualitätsprojekten zeigt, dass diese sehr hilfreich sein kann. Sie liefert genaue Angaben zum Datenformat und die häufigsten Muster eines Attributs. So lassen sich Felder mit festen Formatregeln wie beispielsweise Telefonnummer oder Artikelnummer (EAN-Code) auf ihre Richtigkeit hin prüfen. Auf Basis der eruierten Formate lassen sich dann Standardisierungen der Daten aufbauen, um beispielsweise Daten aus unterschiedlichen Datenquellen oder länderspezifischen Formatregeln in ein gemeinsames und einheitliches Format zu überführen.

Eine Gemeinsamkeit aller Ergebnisreports der unterschiedlichen Ana-

lysen ist der Drill-Down-Bereich. Mit seiner Hilfe kann aus der Grobanalyse der Attribute auf die einzelnen Wertegruppen bis auf die einzelnen Datensätze geschlossen werden. Dies sollte jedoch nur in Stichproben erfolgen, schließlich verfügt so manches Data Warehouse über mehrere Millionen Datensätze in einer Tabelle.

Nach erfolgreichem Profiling lassen sich aus den Analysen Datenregeln ableiten, die sich dann in sogenannten „Korrektur-Mappings“ verwenden lassen. Dieser Begriff wird durch das Tool vorgegeben und kann leider leicht missverstanden werden. Die resultierenden Mappings lassen keine vollständig automatisierte Datenkorrektur zu, sondern sortieren die Daten lediglich nach dem Aschenputtel-Prinzip „Die Guten ins Töpfchen, die Schlechten ins Kröpfchen“ in unterschiedliche Tabellen zur Weiterverarbeitung.

Fazit

Die vorgestellten Analysen lassen sich zwar auch mit eigenen Implementierungen auf SQL-Ebene lösen, der Aufwand für selbstgeschneiderte Alternativen steht jedoch in keinem wirtschaftlichen Verhältnis zu einem toolunterstützten Profiling. Wichtig ist, dass das Profiling als Prozessschritt etabliert und in einen ganzheitlichen Data-Quality-Prozess integriert wird, damit eine kontinuierliche Betrachtung der Daten möglich ist. Ein weiterer entscheidender Punkt für ein erfolgreiches Profiling ist die Auswahl eines erfahrenen Teams, das umfassende Kenntnisse über die Daten sowie Erfahrung mit der Ergebnisinterpretation vorweisen kann.

Weiterführende Literatur

[1] Gartner BI Summit Survey EMEA, Fall 2007
[2] Apel, Brehme, Eberlein, Merighi: Datenqualität erfolgreich steuern, HANSER 2009
[3] Kimball, Caserta: The Data Warehouse ETL Toolkit, Wiley 2004
[4] Hildebrand, Gebauer, Hinrichs, Mielke: Daten- und Informationsqualität, Vieweg+Teubner 2008

Kontakt:
Stefan Schrickel
stefan.schrickel@opitz-consulting.com

Analysemethode	Beschreibung
Eindeutigkeitsanalyse	Es werden bestehende Eindeutigkeitsregeln (Unique Keys) auf ihre Eindeutigkeit untersucht. Ebenso wird der gesamte Datenbestand der Analyseobjekte nach eindeutigen Schlüsseln kontrolliert
Domänenanalyse	Die Datenfelder werden hinsichtlich immer wiederkehrender Dateninhalte analysiert. Dabei werden häufig gefundene Werte (Domänen) als Ergebnis angezeigt und die statistische Verteilung angegeben
Funktionale Abhängigkeitsanalyse	Diese Analyse untersucht die Dateninhalte unterschiedlicher Felder nach Abhängigkeit. Die Ergebnisse können beispielsweise zur Datenmodellierung nach der 3. Normalform genutzt werden
Referentielle Integritätsanalyse	Hier wird der Datenbestand nach möglichen Fremdschlüsseln zwischen unterschiedlichen Tabellen durchsucht. Als Ergebnis wird unter anderem eine grafische Aufschlüsselung nach Orphan- und Childless-Datensätzen geliefert. Ermittelt werden also Daten, die nur in der abgeleiteten Tabelle existieren (Orphans) und solchen, die nur in der übergeordneten Tabelle vorkommen (Childless)
Musteranalyse	Es wird nach häufigen Mustern in den Daten gesucht. Damit lassen sich unter anderem gewisse Formatvorschriften für die Daten validieren. Diese Analyse wird im Folgendem nochmals mit einem Beispiel erläutert, da sie sich – im Gegensatz zu den anderen Datenanalysen ohne eine Toolunterstützung – nur mit extrem hohem Aufwand abbilden lässt
Datentypanalyse	Der für die Dateninhalte vorherrschende Datentyp wird ermittelt. Die Ergebnisse dieser Analyse können für die spätere Datenmodellierung genutzt werden
Aggregationsanalyse	Hier werden technische Analysen wie MIN, MAX, NULL-Value-Ermittlungen vorgenommen. Diese Analyse bietet häufig einen ersten Einstieg in eine detaillierte Datenanalyse, da bereits erste Auffälligkeiten geliefert werden
Benutzerdefinierte Datenregeln	Hier können eigene Datenregeln definiert und im Rahmen des Data Profilings geprüft werden

Tabelle 1: Die verschiedenen Analysemethoden