Datenqualität - Auf dem Weg in eine genauere Zukunft

Daniel Schulte
viadee Unternehmensberatung
GmbH
Anton-Bruchausen-Str. 8
48147 Münster
daniel.schulte@viadee.de

Zusammenfassung

Mit Themen, wie Big Data, muss man neben der Masse, auch die Qualität der Daten im Auge behalten. Wenn die Datenqualität auf einem schlechten Niveau liegt, sind natürlich auch alle Berichte, Analysen und vor allem die darauf basierenden Entscheidungen fragwürdig.
Neben methodischen Wegen Datenqualität von einem Bauchgefühl hin zu einer quantifizierten Größe zu entwickeln, werden auch spezifische Möglichkeiten und Fallstricke in SAS aufgezeigt, wie man sich diesem Thema nähern kann.

Schlüsselwörter: Datenqualität, Prozessüberwachung, Schnittstellenmanagement, Datenmanagement

1 Motivation

Bei der Verwendung von Daten, egal ob aus „eigenen“ Systemen oder aus einer bereitgestellten Schnittstelle, sollte immer die Frage im Hinterkopf mitschwingen: „Ist das, was hier ankommt auch korrekt?“.
In der Projektpraxis – vor allem in Datawarehouse Projekten – ist die Antwort auf diese Frage oft der Unterschied zwischen einem akzeptierten und vertrauenswürdigem Produkt oder einem System, welchem aufgrund von nicht vorhandener Glaubwürdigkeit, ein Untergang schon bei der Einführung vorhergesagt werden kann.
Es geht vor allem nicht nur um eine technische Vollständigkeit, sondern auch um fachliche Plausibilität und Abgleichbarkeit.

2 Was ist Datenqualität?

Wenn es um Qualität geht – das kennt man aus dem Alltag – ist die Definition dieser oft eine sehr subjektive. Aus technischer Sicht mag es ausreichen, dass alle definierten Regeln für die referenzielle Integrität erfüllt sind. Aus fachlicher Sicht reicht dies natürlich noch nicht aus. Hier müssen die Zahlen plausibel und idealerweise mit anderen Systemen abstimmbar sein.

Beispiel:
Beim Zusammenführen von Daten aus zwei operativen Systemen sind ggf. Vereinheitlichungsschritte durchzuführen, welche schon Probleme produzieren können. Beispielsweise sollen Artikelnummern im Datawarehouse numerisch abgelegt werden. System A liefert allerdings dieses Feld als Zeichenkette. Um diese nun als Zahl einzulesen, findet man oft put/input Kombinationen

```
data sys1;
Artikel_nr="12345678";
run;

data dwh;
set sys1;
Artikel_Nummer=input(put(artikel_nr,$8.),8.);
drop artikel_nr;
run;
```

Solange die Artikelnummern in sys1 bis zu acht Stellen haben, funktioniert der Code ohne Probleme. Erhöht man allerdings in der Tabelle sys1 die Anzahl der Stellen auf z.B. “123456789“ funktioniert der Code weiterhin ohne Probleme und es wird auch kein Fehler oder eine Warnung ausgegeben. In der Zieltabelle landen allerdings weiterhin nur 8 Stellen der Nummer und damit der Wert 12345678. Die darauf ausgeführten Auswertungen liefern natürlich damit ein falsches Bild und ggf. auch falsche Entscheidungen.

Dieses kleine Beispiel (so in einem Kundenprojekt gefunden!) zeigt, wie schnell sich fachliche Anpassungen, in diesem Fall die Erweiterung/ Anpassung, eines Nummernkreises auf die Datenqualität auswirken. Den Betreuern des Systems, in welchem die Änderungen implementiert wurden, war zudem nicht bewusst, dass es als Quelle für ein Datawarehouse genutzt wurde und daher war dies auch nicht in dem Projektrahmen mit betrachtet worden.

3 Technisch Maßnahmen

Bereits auf der technischen Ebene kann viel für die Datenqualität gemacht werden.

3.1 Validierung bei Eingabe

Viele Probleme entstehen bereits bei der Erfassung von Daten. Bei der Erfassung durch den Menschen sind dies:

- Rechtschreibfehler
- Unkenntnis der Anwendung und inhaltliche Zuordnung
- Schlechte GUI
- Unzureichende Unterstützung bei der Eingabe

Oftmals werden allerdings Regeln, welche die Eingabe auf Masken prüfen und validieren sollen, abgeschwächt, um Benutzer nicht zu stark zu bremsen. Regelwerke, welche

aus bisher gemachten Eingaben Vorschläge generieren (z.B. aus einer Postleitzahl direkt nur noch valide Ortsnamen zur Auswahl anbieten), können Anwender unterstützen und Eingabefehler vermindern. Auf der anderen Seite, sind z.B. Neubaugebiete bei solchen Assistenten oft ein Problem.

3.2 Schnittstellenmanagement

Werden Daten aus einem System in einem anderen weiterverarbeitet, sollten dies zunächst alle beteiligten wissen. Das Bewusstsein für den Datenaustausch hilft schon dabei, saubere Informationen zu übertragen und Anpassungen zu kommunizieren. Aber auch beim Austausch können, neben den reinen Nutzdaten, auch Metadaten hinzugezogen werden. Fragestellungen, wie:

- Anzahl der Sätze,
- Summe der Beträge über die Schnittstelle,
- Ablieferung und Abholungszeitstempel,

ermöglichen es, die Lieferung und Abholung der Daten deutlich transparenter zu gestalten. Stimmen z.B. die Anzahl der Sätze oder die Prüfsumme(n) nicht überein, liegt ein Problem vor, welches analysiert werden muss. Oder gibt es abgelieferte Daten, die nie oder evtl. mehrfach abgeholt wurden? Die Konsistenz der Daten ist in beiden Fällen nicht mehr gegeben.
Solche Mechanismen flächendeckend zu etablieren, ist vor allem bei heterogenen und gewachsenen Umgebungen nicht immer leicht aber lohnenswert. Bei einer Implementierung in SAS ist auf jeden Fall darauf zu achten, dass auch parallel die Tabelle zur Protokollierung beschrieben werden kann, z.B. per SAS Share oder einer eingebundenen Datenbank-. Ein normales SAS Dataset ist allerdings ungeeignet, da es für den Schreibvorgang komplett gesperrt ist.

3.3 Datenvalidierung im Batch

Bei der Schnittstellenverarbeitung lassen sich ebenfalls Prüfungen einbinden. Hierfür können reguläre Ausdrücke herangezogen werden. Damit können Zeichenketten auf die Einhaltung bestimmter Regeln geprüft werden.
Beispiele:

Gültige deutsche Postleitzahl (5 stellen je 0-9)

```
if prxmatch('/[0-9][0-9][0-9][0-9][0-9]/',plz) then output;
if prxmatch('/\d\d\d\d\d/',plz) then output;
if prxmatch('/\d{5}/',plz) then output;
```

Namen im Muster „Nachname, Vorname“

```
if prxmatch("/[A-Z]\w+, [A-Z]\w+/",name) then output;
```

4 Fachliche / inhaltliche Maßnahmen

Auch wenn auf technischer Ebene schon aktiv für bessere Datenqualität gearbeitet wurde, sollte auch fachlich inhaltlich validiert werden.

4.1 Werteverteilung

Bei dem unter 2. genannten Beispiel wurde das Problem aufgrund der Werteverteilung identifiziert. In einer Analyse nach Artikelgruppen fiel auf, dass einige Artikel deutlich zu hoch bewertet und keine der neuen Artikelnummern auffindbar waren. Mit solchen Verteilungsanalysen und deren Abgleich mit Zuliefersystemen können Diskrepanzen automatisch identifiziert werden.

4.2 Ableitung von Kennzahlen

Manchmal kann es aber auch notwendig sein, Kennzahlen abzuleiten, um Abweichungen zu identifizieren. In einem anderen Beispiel hat sich folgendes Bild ergeben:

In einer multidimensionalen Analyse und deren Migration auf ein anderes System musste die Ermittlung von Kennzahlen erneut durchgeführt werden, da hier eindeutige IDs je Betrachtungszeitraum gezählt werden sollten, und das alte System nur aggregierte Werte beinhaltete. Bei dieser Werteermittlung konnte bis auf einen kleinen Teil von IDs alles korrekt ermittelt werden. Bei dieser kleinen Menge fiel allerdings auf, dass diese, gemessen am ermittelten Umsatz, einen viel zu hohen pro Kopf Umsatz aufwies. Diese abgeleitete Kennzahl hat zur Identifizierung einer fehlerhaften Implementierung in der ursprünglichen Werteermittlung geführt, welche seit mehreren Jahren als akzeptierte Kennzahl im Unternehmen genutzt wurde.

5 Datenqualität als permanente Aufgabe

Das Problem mit der Datenqualität ist leider oft, dass sie mit der Zeit schlechter wird. Daher kann es sein, dass ein System bei der initialen Abnahme hervorragende Daten liefert. Mit jeder Anpassung, durch „Tricks“ von Endanwendern und alleine durch Alterung, kann sich dieses Bild mit der Zeit leider sehr wandeln. Daher ist die Erhebung der Datenqualität nicht eine einmalige Aktion, sondern sollte interdisziplinär als permanente Aufgabe etabliert werden. Bei dieser Aufgabe muss neben dem technischen Know-how auch Wissen über die Fachlichkeit vorhanden sein. In der Regel ergibt sich daraus ein Team aus IT- und Fachabteilungen. Für das hierfür nötige Budget ist die Einbindung der Geschäftsführung als „Sponsor“ unabdingbar. Leider lassen sich fehlerhafte Daten initial nur in den seltensten Fällen mit einem monetären Wert versehen. Aber in immer mehr vernetzen Systemen hat ein kleiner Fehler ggf. größere Auswirkungen und kann eine geschäftsrelevante Entscheidung in die falsche Richtung laufen lassen. Die Maßnahmen dürfen allerdings nicht nur als „Feuerwehreinsätze“ wahrgenommen werden sondern sollten die Probleme an der Wurzel angehen.

5.1 Definition von „Qualität“

Um eine vergleichbare Aussage zur Datenqualität liefern zu können, muss als Teilaufgabe definiert werden, in welchen Metriken analysiert und bewertet werden soll. Hierzu kann z.B. gelten:

- Sind alle Schlüsselfelder valide belegt
  - Wenn nein => Fehler
- Sind Artikel gültigen Gruppen zugeordnet
  - Wenn mehr als 10 Artikel keine Zuordnung haben => Fehler
- Sind Warentransaktionen im Logistiksystem enthalten
  - Sind mehr als 0,5% der Transaktionen nicht enthalten => Fehler

Solche Metriken können vielschichtig sein und auch abgestuft werden. Also muss nicht zwischen Schwarz / Weiß unterschieden werden, sondern akzeptable Abweichungen, welche z.B. durch unterschiedliche Datenlieferzeitpunkte entstehen, können hingenommen werden. Manche Daten werden, beispielsweise aufgrund von Laufzeiten, nur monatlich aktualisiert und daher sind bestimmte Abhängigkeiten bewusst noch nicht erfüllbar.

5.2 Messen

Die aufgestellten Metriken müssen auch erfasst werden. Leider passiert es zu Hochlastzeiten, wie einem Monatswechsel, oft, dass Ressourcen knapp werden und die Erfassung der Datenqualität wird dann gerne „Opfer“ von spontanen Optimierungen.

5.3 Analysieren

Sind die Daten erfasst, müssen sie natürlich analysiert und interpretiert werden. Nur so lassen sich die relevanten Maßnahmen für die Verbesserung definieren und mögliche Seiteneffekte identifizieren.

5.4 Verbessern

Eine schnelle Verbesserung kann als Ad-hoc-Maßnahme platziert werden. Daten können in der ETL Strecke „optimiert“ oder im Quellsystem manuell angepasst werden. Langfristig hilft es nur den Pfad der Daten zur Quelle zurück zu wandern und dort den Fehler zu beseitigen. Allerdings muss hier abgewägt werden, ob eine Anpassung im Quellsystem wirtschaftlich ist oder ob die möglichen Seiteneffekte durch Anpassungen ggf. mehr zerstören als nutzen.

5.5 Erfolgskontrolle

Nach der Implementierung der Verbesserungen sollte dies in der Erfolgskontrolle sichtbar werden. Über ein Dashboard können z.B. die zugrundeliegenden Metriken veröffentlicht werden und damit auch den gesamten Prozess transparent gestalten.

6 Verantwortlichkeit

Da es sich bei der Analyse von Datenfehlern auch um die Analyse von Arbeitsergebnissen von Kollegen handelt, hat man ggf. nicht nur technische oder fachliche, sondern auch soziale Hindernisse, um die man sich kümmern muss.
Bei der Spurensuche werden ggf. fehlerhafte Implementierung auf Seiten der IT aufgedeckt oder Fachkonzepte bzw. Geschäftsprozesse, welche unzulässig sind. Die Analyse und die Lösungsstrategie sollte daher immer ziel- und nicht problemorientiert sein. Wenn ein Fehler gefunden wurde, liegt der Fokus auf der Behebung und nicht, auf denjenigen, der es genau verursacht hat.
An dieser Stelle hat es sich bewährt die Datenqualitätsanalyse mit externer Unterstützung durchzuführen. Hier gibt es geringere zwischenmenschliche Hürden und man kann Probleme schneller aus der Welt schaffen.

7 Gründe für das Scheitern

Leider kann man auch ein paar Ecken feststellen, an denen Maßnahmen zur Steigerung der Datenqualität scheitern.

7.1 Kein kontinuierlicher Prozess

Nach schnellen Erfolgen wird der Prozess oftmals wieder eingestellt, da man ja die größten Probleme beseitigt hat. Da diese mit der Zeit aber wieder anwachsen, gehen zwischenzeitlich Know-how und Vergleichsmöglichkeiten verloren und müssen kostenintensiv wieder aufgebaut werden.

7.2 Keine Vorgaben der Fachabteilung / keine Organisation des Prozesses

Ohne diese Vorgaben lassen sich nur schwer sinnvolle Metriken definieren. Im Rahmen ihrer Aufgaben brauchen Mitarbeiter aus IT- und Fachabteilung auch den Freiraum, diesen nachzukommen. Dies kann nicht als Nebentätigkeit mit dem normalen Tagesgeschäft erledigt werden. Eigenes Budget und eine Freistellung für diese Aufgabe ermöglichen es den Mitarbeitern sich deutlich besser in das Thema einzubringen.

7.3 Beschränkung auf Datenmigration

Sollte die Handlungsreichweite rein auf die Datenmigration beschränkt sein, kann es schnell zu unlösbaren Problemen kommen. Immer wieder Sonderregeln in ETL Prozesse zu implementieren, kann keine Dauerlösung sein. Mit steigender Anzahl der Systeme entsteht schnell ein Zoo von Sonderregeln, der ein performantes Arbeiten oder eine Wartung unmöglich macht.

8 Fazit

Datenqualität wird leider oft noch nicht als Geschäftsvorteil gesehen. Mit steigender Integration von Systemen in der Entscheidungsfindung und dem Ankauf von Daten wird die Bewertung immer wichtiger. Spätestens, wenn Mitbewerber durch effektivere Kundenansprachen und bessere Erkenntnisse aus den eigenen Daten mehr Umsatz generieren, wird es entscheidend, sich um einen entsprechenden Prozess zu kümmern.