



Nicht nur Software zählt

Voraussetzungen für erfolgreiche Datenqualitätsprojekte

Ohne Datenqualität bleiben BI und CRM erfolglos. Um zu verlässlichen Daten zu gelangen, müssen Unternehmen leistungsfähige Software-Produkte zur Datenbereinigung einsetzen, doch damit alleine ist das Problem nicht zu lösen. Erst die richtige Mischung von Methode und Technologie sichert den Erfolg von Datenqualitätsmaßnahmen. Praxisbewährte und akademisch fundierte Methoden sind eine notwendige Voraussetzung für den erfolgreichen und nachhaltigen Einsatz modernster Such- und Bereinigungsverfahren im Datenqualitätsmanagement.

Daten wie Produkte managen

Zwischen Produkten und Daten besteht eine bemerkenswerte Analogie: Daten werden wie Produkte erzeugt, gekauft, gelagert, normiert, veredelt, benutzt und letztendlich ausgeliefert. Sie bestimmen in ähnlich großem Maße die Abläufe in den Unternehmen.



Durch Neuerfassung werden Daten wie z. B. ein Lieferschein erzeugt. Für ein Mailing werden Adressen gekauft. Daten werden in Datenbanken oder Archiven gelagert bzw. gespeichert. Auch für Daten gilt, dass sie bestimmten Normen entsprechen sollten, die firmenspezifisch sein können, z. B. Datumsformat „TT.MM.JJJJ“. Durch eine Warenkorbanalyse lassen sich die Daten aus den Scannerkassen einer Supermarktkette veredeln zu Erkenntnissen über das Einkaufsverhalten der Kunden. Referenzdaten wie eine Liste aller Bankleitzahlen und Institutsnamen können genutzt werden, um die Bankverbindungsdaten korrekt zu erfassen. Auch Daten werden ausgeliefert, sei es intern an die Marketingabteilung zur Durchführung einer Kampagne oder die Übermittlung von Lieferadressen an die mit der Auslieferung beauftragte Spedition.

Ähnlich wie Produkte haben auch Daten einen Produktionsprozess, der aber nicht explizit beschrieben und – in den wenigsten Fällen – optimiert ist. Dies liegt daran, dass es keinen Produktmanager für Daten gibt. Die Bedeutung der Daten für die Steuerung betrieblicher Abläufe legt nahe, Daten wie Produkte zu managen. Dafür ist es unerlässlich, die Bedürfnisse der Kunden, d. h. der

Datennutzer zu verstehen, um den Daten-Produktionsprozess entsprechend optimieren zu können. Analog zu Produkten gilt auch hier, dass sich dauerhaft hohe Datenqualität nur gewährleisten lässt durch ein Qualitätsmanagement, das den gesamten Lebenszyklus der Daten abdeckt.

Bewusstsein für Datenwert schaffen

Im Daten-Produktionsprozess spielen die Mitarbeiter die wesentliche Rolle bei der Pflege der Daten. Datenerfasser (data collector) erzeugen neue Daten und bringen sie somit in die Systeme. Sie verändern die Daten oder aktualisieren sie. Die Datennutzer (data consumer) sind die Abnehmer der Daten. Sie benötigen die Daten in einer bestimmten Qualität, um ihre Aufgaben erledigen zu können.

Die Datennutzer sind der eigentliche Grund dafür, dass bestimmte Daten überhaupt erfasst und gespeichert werden.

Zwischen data consumer und data collector steht der Datenverwalter (data custodian). Er ist zuständig für den Betrieb der Anwendung, Erteilung von Zugriffsrechten, Backup, usw. und sorgt somit dafür, dass alle Berechtigten jederzeit Zugriff auf die Daten haben.

Um Abhängigkeiten und Auswirkungen fehlerhafter Daten allen Beteiligten klar zu machen, ist es wichtig, dass sich jeder Mitarbeiter über seine Rolle in Bezug auf die Daten bewusst ist. Dem data consumer muss als Abnehmer der Daten bewusst sein, dass er seine Anforderung an Datenqualität deutlich formulieren muss, genauso wie sich der data collector bewusst sein muss, welche Auswirkungen Datenqualitätsmängel in anderen Unternehmensbereichen haben. Nur wenn alle im Bewusstsein ihrer Rolle am Datenproduktionsprozess arbeiten, kann Datenqualitätsmanagement erfolgreich sein.

Datenqualität messen

Auch für Datenqualität gilt: Man kann nur das verbessern, was man auch messen kann. Daher steht am Anfang eines erfolgreichen Datenmanagements die Einführung von zuverlässigen Messverfahren. Diese Messverfahren müssen aussagekräftige Kennzahlen zur aktuellen Datenqualität liefern und sich jederzeit wiederholen lassen. Nur so lassen sich Veränderungen der Datenqualität zuverlässig feststellen. Die Berechnung von Datenqualitätskennzahlen geschieht mit Hilfe von Metriken, die auf Datenqualitätsmesswerte angewandt werden. Ein Datenqualitätsmesswert ist die Anzahl von Datensätzen, die eine Datenqualitätsregel verletzen. Z. B. die Anzahl der Datensätze, deren Wert im Feld „Postleitzahl“ nicht aus 5 Ziffern besteht. Die zu Grunde liegende Datenqualitätsregel wäre in diesem Fall:

Deutsche Postleitzahlen sind fünfstellig numerisch.



Datenqualitätsregeln beschreiben die Geschäftsregeln (business rules), denen Daten gehorchen müssen, um von hoher Qualität zu sein. Ein Teil dieser Regeln lässt sich aus gesetzlichen Bestimmungen, offiziellen Standards, Firmenrichtlinien und Arbeitsanweisungen ableiten. Ein großer Teil dieser Regeln lässt sich nicht nachschlagen. Diese Regeln sind teilweise durch die Gestaltung der Eingabemasken vorgegeben oder existieren nur in den Köpfen ihrer Mitarbeiter. Gerade diese Regeln müssen entdeckt und durch Dokumentation explizit sichtbar gemacht werden.

Vorgehensmodell für nachhaltig hohe Datenqualität

Datenqualitätsprojekte gelingen dann, wenn Methoden und Software-Technologien abgestimmt eingesetzt werden: Es müssen zunächst die anwenderspezifischen Ziele genau definiert und die Ausgangssituation analysiert werden. Auf dieser Basis können dann die erforderlichen technischen, prozessualen und kulturellen Voraussetzungen geschaffen werden, damit moderne Software-Technologien erfolgreich eingesetzt werden können. Mit dem FUZZY! DataCare Process existiert ein bewährtes Vorgehensmodell, das durch eine systematische Situationsanalyse, gezielte Prozessoptimierung und den koordinierten Einsatz von Datenqualitätsstools den Erfolg von Datenqualitätsprojekten garantiert.

Methoden des Datenqualitätsmanagement

Punktuelle Datenbereinigungen, die häufig ereignisgetrieben durchgeführt werden, erzielen nur kurzfristige Wirkung. Eine wirklich effektive Datenqualitätssteigerung, die auch nachhaltig den Erfolg eines Unternehmens sichert, basiert auf den Techniken des Total Data Quality Management (TDQM). Dabei werden die im Unternehmen vorhandenen Daten entlang ihres gesamten Lebenszyklus (Data Lifecycle) analysiert und qualitätsgesichert.

Der FUZZY! DataCare Process bietet auf Grundlage des TDQM einen kompletten Projektrahmen zur erfolgreichen Abwicklung von Datenqualitätsprojekten.

Dabei werden zunächst in einer ersten Ist-Aufnahme die Daten ausgewertet und gemeinsam mit dem Kunden wird die Definition der individuellen Datenqualitätsregeln durchgeführt. Diese Datenqualitätsregeln spiegeln die Anforderungen der Unternehmen an die Datenqualität wider. Darin finden sich Geschäftsprozessregeln des Unternehmens aber auch standardisierte oder rechtliche Vorgaben für Informationen wieder.

Nach der Analyse der Datenfehler gilt es, im darauffolgenden Schritt die Ursachen zu bekämpfen und für die Verbesserung der Datenqualität zu sorgen. Hierbei können einmalige Bereinigungen für die initiale Optimierung sorgen. Zugleich gilt es aber, permanente qualitätssichernde Maßnahmen, zum Beispiel durch Prozessoptimierung oder Integration von qualitätssichernden Werkzeugen direkt in die Anwendungen, zu ergreifen. Durch eine fortlaufende Messung der Datenqualität als Monitoring-Prozess kann die aktuelle Datenqualität permanent überwacht werden.

Joachim Schmid

Leiter Software-Entwicklung

FUZZY! Informatik AG

E-Mail: joachim.schmid@fazi.de



Mit Hilfe dieser Datenqualitätsregeln erfolgt dann die eigentliche Messung der Datenqualität. Die gemeinsam festgelegten Regeln werden auf den gesamten Datenbestand angewandt. Es wird ermittelt, welche Datensätze diesen Regeln widersprechen. Die statistische Auswertung dieser Messung stellt eine Gesamtgüte, den Datenqualitätsindex, bereit.

In der darauf folgenden Analyse werden die Ursachen für die in der Messung ermittelten Datenfehler identifiziert. Diese können sowohl innerhalb der Daten stecken, als auch in den Prozessen, die die Daten übertragen, oder in den Anwendern selbst.