Fachartikel

Unterstützung der Datenqualitätsverbesserung im Rahmen der Data Governance

von

Hon.-Prof. Dr. Leonhardt Wohlschlager, CGI[1]

[1] An dieser Stelle möchte ich meinen französischen Kollegen für unsere gute Zusammenarbeit und ihre exzellenten Inputs zu diesem Papier herzlich danken, namentlich insbesondere den Herren Julien Bourgois, Arnaud Deneux und Xavier Landry.

Management Summary

Im Rahmen einer IT Governance ist der Teilbereich der Data Governance die Fähigkeit eines Unternehmens, einen Geschäftsprozess zur Datenqualitätsverbesserung in einem Projekt einzuführen, ihn zu automatisieren und kontinuierlich zu verbessern, um mit diesem Prozess die Qualität der Daten ebenfalls zu steigern. Die Motivation für die Umsetzung einer Data Governance kann auf der Umsetzung und Einhaltung einer bestimmten Datenqualität aus unternehmerischen Zielvorgaben (Governance), aus regulatorischen, vertraglichen und gesetzlichen Vorgaben (Compliance) und/oder aus Vorgaben des Risikomanagements (Security) beruhen.

Die vorliegende Publikation beschreibt die eher technischen Aspekte der Data Governance und liefert Praxisbeispiele aus der Versicherung zur Profilierung (Data Profiling), Prüfung (Data Auditing), Korrektur und Bereinigung (Data Cleansing), Veredelung (Data Augmentation) und Überwachung der Daten (Data Monitoring) als Grundlage für die Unterstützung des Data-Governance-Prozesses.

Mit Hilfe dieser Schlüsselkonzepte werden die Unterstützungspotenziale von Data-Governance-Werkzeugen identifiziert und erklärt, so dass vor allem die Führungskräfte eine Grundlage für ihre Softwareauswahlentscheidung erhalten.

Diese Publikation wurde erstellt für Führungskräfte und Spezialisten, die in Fach- und IT-Bereichen für die betrieblichen Daten, ihre Qualität und ihr Management verantwortlich sind. Vorschläge zu konkreten Softwareanwendungen werden hier nicht diskutiert, da jedes Unternehmen individuelle Anforderungen und Voraussetzungen hat.

Kreislauf des Datenqualitätsmanagements

Die folgende Abbildung beschreibt den Zyklus verschiedener methodischer Einzelaktivitäten zur Verbesserung der Datenqualität im Rahmen der Data Governance in ihrem Zusammenhang:

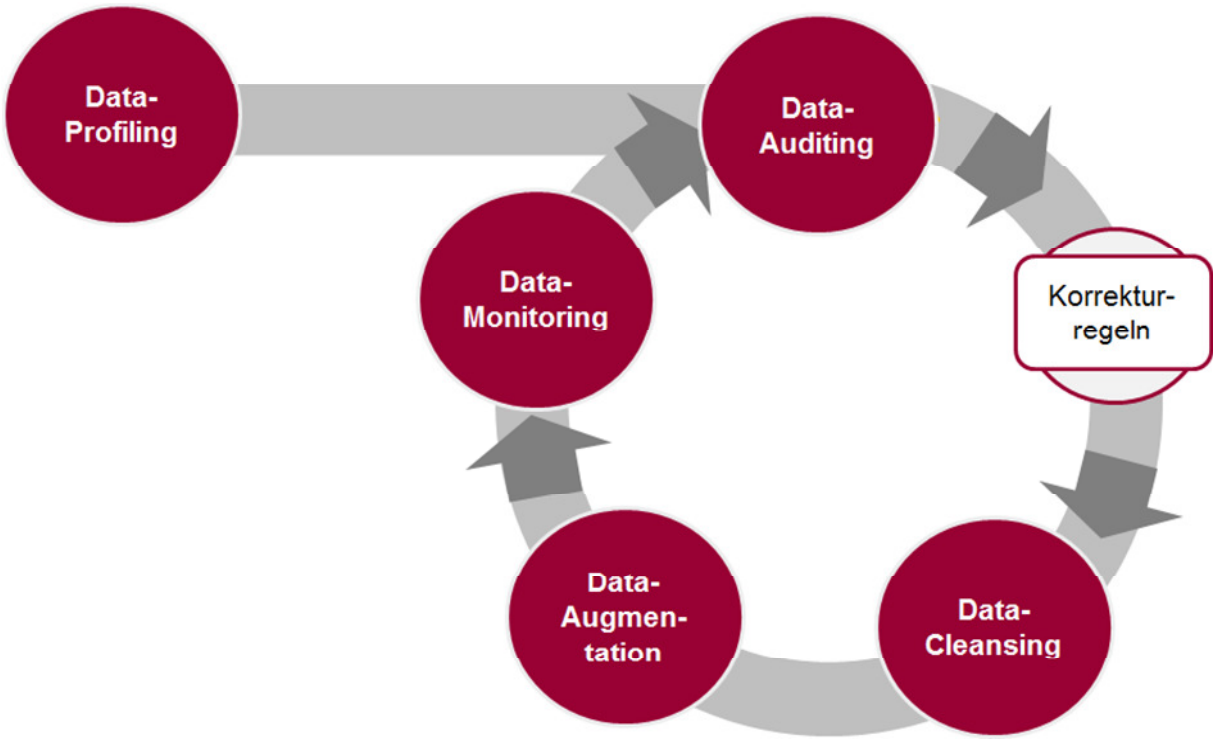

Abbildung 1: Zyklus des Datenqualitätsmanagements

Mit dem Data Profiling erfolgt zunächst eine statistische Analyse, die mit dem Data Auditing durch eine Business-Analyse vervollständigt wird. Dabei werden Datenqualitätskriterien wie z.B. die Vollständigkeit, Fehlerfreiheit und/oder Angemessenheit der Daten geprüft sowie Geschäftsregeln abgeleitet, mit denen die Korrekturen erfolgen. Hiernach erfolgt eine Definition des Objekts, auf das die identifizierten Korrekturregeln angewendet werden. Mit dem Data Cleansing erfolgt schließlich die Korrektur, bei der Dubletten beseitigt sowie Standardisierungen und Bereinigungen durchgeführt werden. Im Anschluss sorgt die sog. Data Augmentation für eine Veredelung (d.h. Anreicherung und Validierung) der Daten. Das Ende des Kreislaufs besteht in dem Data Monitoring, das den Nutzen der durchgeführten Datenkorrektur- und Veredelungsaktivitäten misst und überwacht.

Data Profiling

Durch Data Profiling können Fehler in den Daten identifiziert und letztlich beseitigt werden. Das Data Profiling besteht in einer statistischen Analyse der Daten, die vor allem auf die Qualitätskriterien der Genauigkeit und Eindeutigkeit abzielt (siehe den Glossar im Anhang). Dabei sollen die Charakteristika in den Daten entdeckt und durch das Business erklärt und interpretiert werden. Danach muss das Business die Geschäftsregeln definieren, welche die Daten einhalten sollen. Analyseobjekte können Attribute (Spalten), Tabellen und/oder Kreuztabellen sein. Bei Attributen werden Charakteristika wie

z.B. vorkommende Typen, Minimal-/Maximalwerte, die Eindeutigkeit (Nullwerte, leere Werte, unterscheidbare Werte), die Datenmaske („Patterns“) und der Wertbereich mit Häufigkeitsverteilung analysiert (vgl. z.B. Abbildung 2). Bei Tabellen werden die funktionalen Abhängigkeiten zwischen den Spalten analysiert (z.B. Suche nach Duplikaten, Entdeckung der Business-Identifier für ein Zieldatenmodell). Bei Kreuztabellen werden die referentiellen Beziehungen zwischen verschiedenen Tabellen analysiert, wie z.B. die Analyse der Fremdschlüssel, der Konfiguration (z.B. Orphan-Tabellen) sowie der Nachweis redundanter Daten. Dies kann in den Bau eines normalisierten relationalen Zieldatenmodells sowie in die Erstellung eines Metadaten-Dictionary münden.

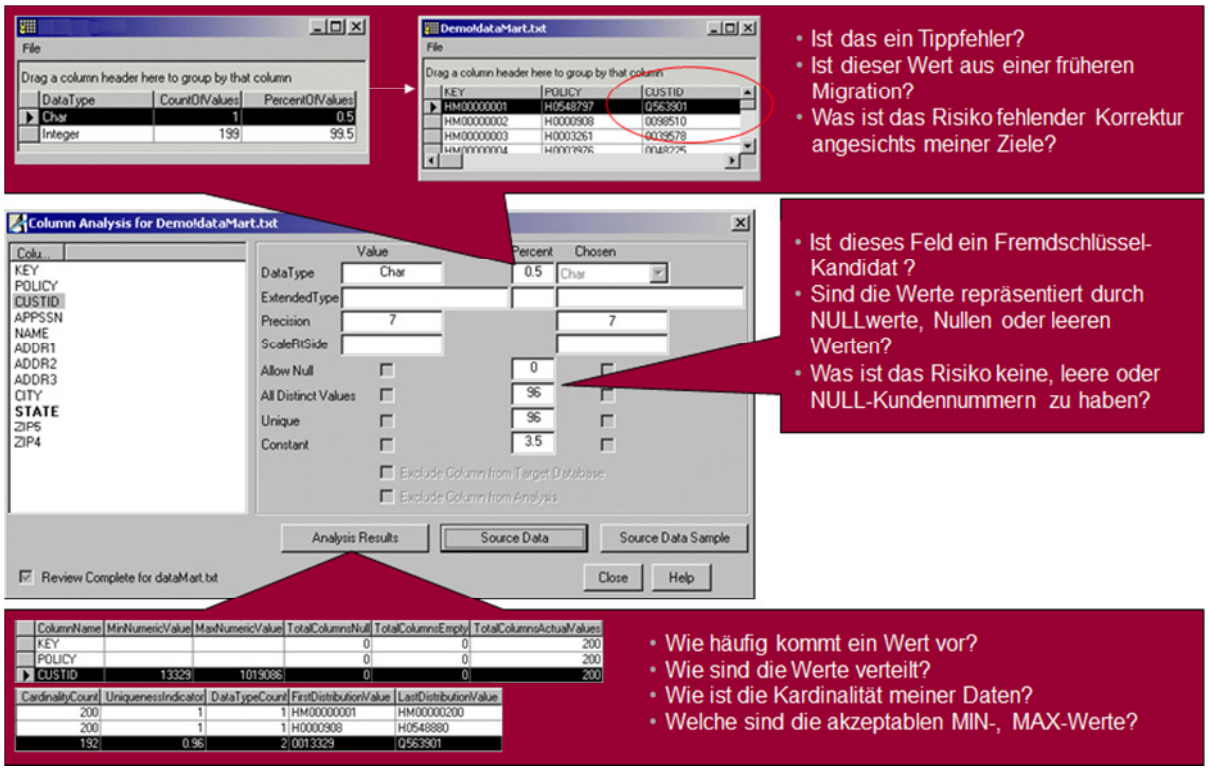

Abbildung 2: Data Profiling-Beispiel "Spaltenanalyse“ (Projektbeispiel)

Die Abbildung 2 oben illustriert die wesentlichen Analysefragen bei der Spaltenanalyse. Die Abbildung 3 unten zeigt ein Profiling-Beispiel für den versicherten Wert bzw. die Versicherungssumme in der Hausrat-/Wohngebäudeversicherung.

Analysiertes Feld		Versicherter Wert oder Versicherungssumme		
Name des Feldes		MVNIM		
Anzahl Datensätze		2.458.650		
Anzahl der eingegebenen Werte	In %	2.458.650		100%
Anzahl der nicht eingegebenen Werte	In %	0		0%
Min/Max/Median Wert		0	3.663.530,23	0
Gesamtbetrag der Versicherungssummen		288.712.116,32		
Test durchgeführt unter [anonymisiert]		n/a		
Anzahl der positiven Testergebnisse	In %	n/a		In %
Anzahl der negativen Testergebnisse	In %	n/a		In %

Verteilung :
Liste der Extremwerte :

Feldwert	Anzahl
0	2.420.321
0,15	3
0,3	69
1,07	1
1,52	4

Feldwert	Anzahl
3.663.530,23	1
1.531.367,76	1
1.481.108,0	1
1.435.265,27	1
1.401.446,0	1

Abbildung 3: Data Profiling-Beispiel "Versicherter Wert/Versicherungssumme"[2]

Das Feld enthält in 100% aller Fälle einen numerischen Wert, hat aber in 2.420.321 (98,44%) Fällen einen Wert von Null.

Weitere Qualitätsprüfungen bestehen z.B. darin festzustellen, ob

- der Buchungscode einer Versicherungspolice mit einem Eintrag in den Büchern korrespondiert (technische Prüfung)
- das Geburtsdatum eines Kunden vor seinem Versicherungsvertragsabschluss liegt (funktionale Prüfung).

Data Auditing

Das Data Auditing erfolgt nach dem Data Profiling und besteht in einer Business-Analyse der Daten, die vor allem auf Business-Anforderungen wie z.B. die Integrität und Konsistenz abzielen. Ziel des Data Auditing ist ein funktionale, unabhängige Prüfung der Daten, die mit Hilfe von zu erstellenden Analyseberichten und elementaren Dashboards (vgl. z.B. Abbildung 4) ermittelt, ob die Datenqualität

[2] Zur Effizienzerhöhung gegenüber konventionellen Abfragewerkzeugen wird Data-Profiling-Software eingesetzt.

eingehalten wird. Eine Folgeaktivität besteht in der Definition von Geschäftsregeln, die auf die Daten angewendet werden, um die Qualität in Bezug auf funktionale Regeln (z.B. gesetzliche Vorschriften) zu korrigieren.

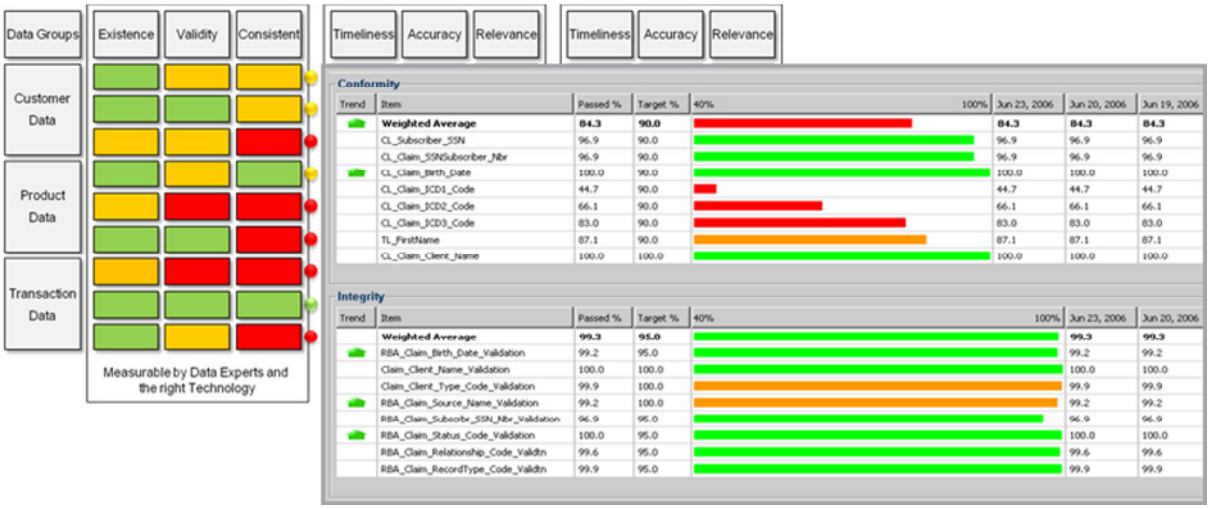

Abbildung 4: Data-Quality-Scorecard[3]

Die Abbildung zeigt im Hintergrund eine Data-Quality-Scorecard, die den verschiedenen Datengruppen auf fachlicher Ebene je nach Qualitätskriterium, z.B. Vollständigkeit, einen Farbcode zuordnet (Colour Coding). Die gelben und grünen Farbcodes zeigen z.B. eine mittlere bis hohe Vollständigkeit an. Der Farbcode ergibt sich aus verschiedenen vordefinierten „technischen" Schwellwerten, deren über- oder unterschreiten den entsprechenden Ampel-Code in Form eines Dashboards auf technischer Ebene hervorbringt. Im Vordergrund findet man das Dashboard einer Software, welche die fachliche Data-Quality-Scorecard durch ein Colour Coding auf technischer Ebene mit größerem Detail unterstützt.

Qualitätsprüfungen bestehen z.B. darin festzustellen, ob

- die Zahlweise der Prämie konsistent mit dem Vertragstyp ist (Business-Konsistenz-Prüfung)
- die in ein Risk-Data-Mart importierten Vertragsdaten mit den aggregierten Daten des Hauptbuchs übereinstimmen („Hauptbuch-Abstimmung").

Data Cleansing

Das Data Cleansing zielt auf die Korrektur der Daten mittels der identifizierten Kriterien ab. Dabei werden die Daten bereinigt, standardisiert und/oder etwaige Dubletten beseitigt. Im Fall der Datenstandisierung werden z.B. Freiformfelder so strukturiert (Parsing), dass Zeichenketten mit mehreren Domänen in mehrere Felder geschnitten und jeweils einer Domäne zugeordnet werden (vgl. Abbildung 5):

[3] Vgl. CGI (Hrsg.) /BI Framework 2009/ 132.

Police
4711 gem. Leben $ 120000
4712 fondsgebundene LV USD 1800
0001232332 Kapital EUR 324

ID	Vertragstyp	Währung	Jahresprämie
4711	gem. Leben	$	120000
4712	fondsgebun-dene LV	USD	1800
0001232332	Kapital	EUR	324

Kunde
Herr Frank Eversz 02/02/1960
Frau Eva Mühlens 08/14/67

Anrede	Vorname	Name	Geburtsdatum
Hr.	Frank	EVERSZ	02/02/1960
Fr.	Eva	MÜHLENS	14/08/1967

Abbildung 5: Data Cleansing-Beispiel: Datenstandardisierung durch Parsing und Domänenzuordnung

Die Homogenisierung der Daten erfolgt dabei auf der Grundlage der durch die Fachabteilungen identifizierten Regeln unter Ergänzung der lokal geltenden Standards. Auf diese Weise können wichtige Daten aus Freiformfeldern gewonnen werden.

Bei der Dublettenbeseitigung, die auf das Datenqualitätskriterium der Eindeutigkeit abzielt, werden zunächst Schlüssel, Gruppierungsregeln und explizite Regeln zur Identifikation des richtigen Datensatzes definiert. Im Anschluss daran können die Dubletten identifiziert und beseitigt werden, wonach eine Prüfung der Ergebnisse erfolgt.

Beispielsweise kann zur Beseitigung von Firmenkunden-Dubletten im Firmenkundengeschäft zunächst eine Attributkombination „Firmenname, Adresse, ID Code“ als Schlüssel zur Identifikation der Firmenkunden definiert werden. Dieser Schlüssel erlaubt den Vergleich zwischen deterministischen Datensätzen und bestimmt den Grad der Ähnlichkeit dieser Datensätze.

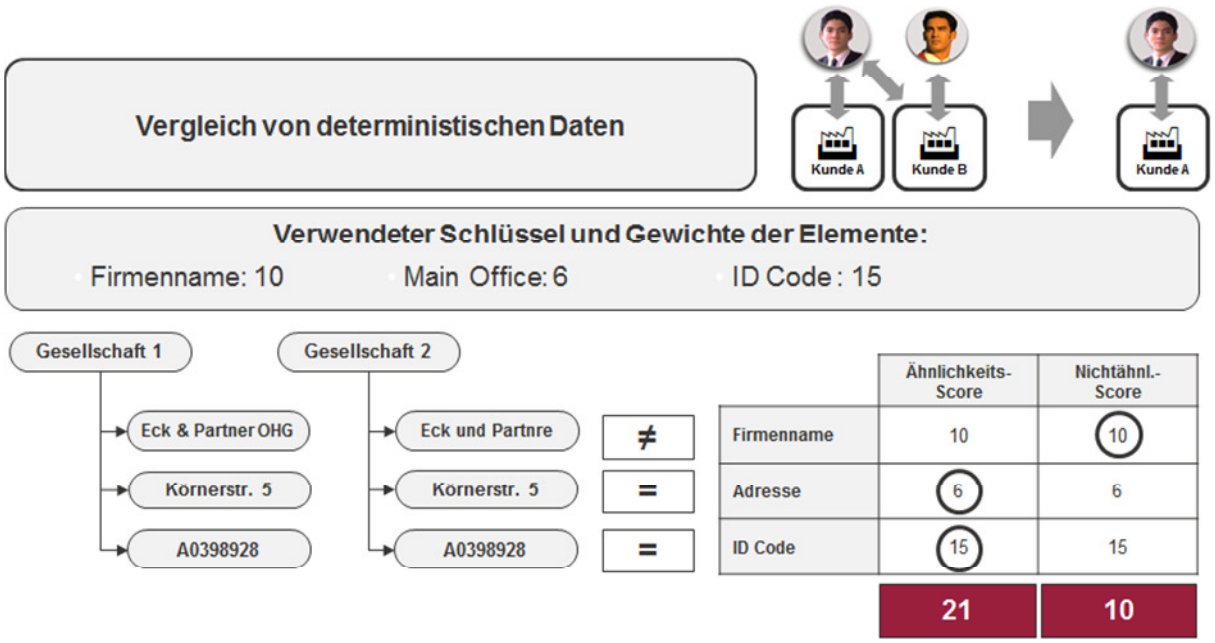

Abbildung 6: Vergleich deterministischer Daten zur Dublettenidentifikation

In der Abbildung 6 wird jedem Schlüsselattribut ein Punktwert (Score) zugeordnet, der sich aus seiner Gewichtung ergibt. Wenn z.B. als eingegebene Adresse wegen eines Eingabefehlers «Eck und Partner» als Wert vorliegt, beträgt der Ähnlichkeits-Score 21 und der Score der Nicht-Ähnlichkeit 10. Auf diese Weise kann die Ähnlichkeit zweier oder mehrerer Datensätze ermittelt werden. Der nächste Schritt besteht in der Identifikation desjenigen Datensatzes, der die Realität am besten abbildet. Dies kann ebenfalls mittels Geschäftsregeln geschehen. Alle übrigen Datensätze können gelöscht werden. Alternativ können Datensätze mittels Analyse- und Verdichtungsmethoden zur Datenqualitätsverbesserung, z.B. der Predictive-Analysis-Methode, automatisch zusammengefasst und optimiert werden.

Data Augmentation

Die sog. Data Augmentation zielt auf die weitere Verbesserung der Daten nach ihrer initiellen Korrektur ab. Dabei werden die Daten angereichert und validiert (vgl. Abbildung 7). Auf Basis dedizierter, z.B. phonetischer, Algorithmen werden die Daten einerseits geprüft und ggf. gelöscht. Alternativ werden sie zur Steigerung ihrer Relevanz mit Daten von Dritten ergänzt. Die Prüfung umfasst z.B. auch die Konsolidierung manueller Dubletten.

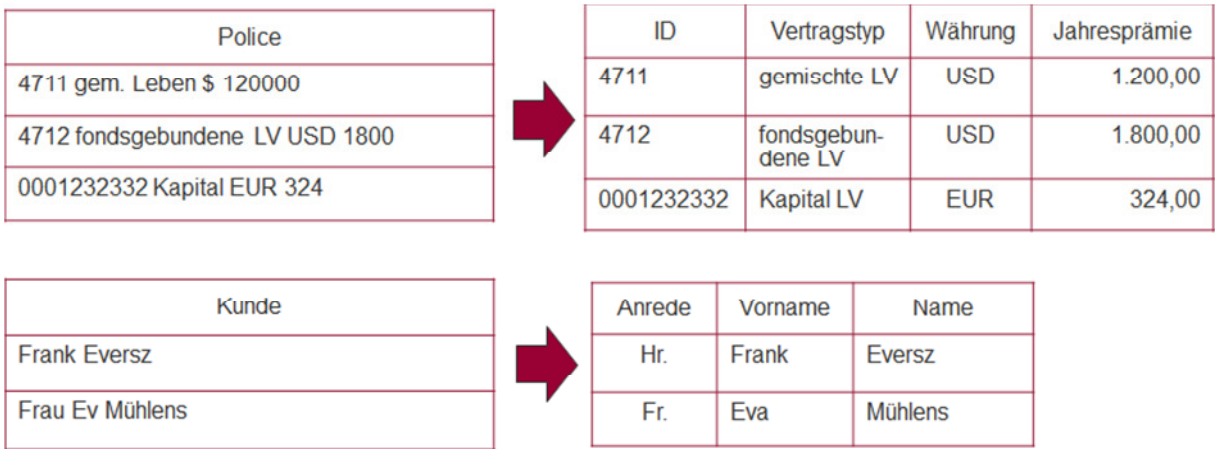

Abbildung 7: Beispiel zur Data Augmentation

Ein Beispiel für eine Data Augmentation ist die Korrektur von Datensätzen, die Jahresbruttoprämien mit fehlendem Komma enthalten, um zwei Zehnerpotenzen.

Zur Erfüllung regulatorischer Vorgaben muss häufig jede Korrektur an den Daten gerechtfertigt und nachvollziehbar dokumentiert werden.[4] Wegen der Nachvollziehbarkeit dürfen die Rohdaten auch nicht überschrieben werden. Stattdessen müssen die angepassten Daten und durch Experten geschätzten Näherungswerte, zusammen mit den Verarbeitungsschritten, neben den unveränderten Rohdaten gespeichert werden. Um die Effizienz der Validierung zu erhöhen, müssen die Korrekturregeln automatisiert und historisiert werden.

[4] Vgl. zur Nachvollziehbarkeitsanforderung in Solvency II z.B. EIOPA (Hrsg.) /Technical Provisions (CP43) 2009/21.

Data Monitoring

Im Data Monitoring schließlich überwachen die Data-Stewards die Datenqualität. Eine permanente, d.h. tägliche Überwachung ist nötig, weil sich die Datenqualität laut Expertenschätzung um durchschnittlich 2% pro Monat verschlechtert. Dabei stellen die Data-Stewards fest, wann und wie sich die Datenqualität ändert. Das Data Monitoring umfasst ebenfalls die Überwachung der Performance der aufgesetzten IT-Systeme, basierend auf objektiven Data-Quality-Key-Performance-Indikatoren wie auch subjektives Expertenurteil.[5] Die Voraussetzung für das Data Monitoring ist eine Bereitstellung dedizierter Mittel.

[5] Vgl. die Kommentare zum Solvency II Data Monitoring in EIOPA (Hrsg.) /Summary Comments 2009/ 47 und 71 ff.

Ausblick

Governance, Compliance und Risikomanagement stellen neue Anforderungen an die Datenqualität. Die Umsetzung der Compliance, die Reduzierung von Bußgeldern und die Erzielung von Wettbewerbsvorteilen sind die wirtschaftlichen Argumente für die Einführung neuer Software zur Unterstützung der Datenqualitätsverbesserung im Rahmen der Data Governance.

Diese Publikation hat verschiedene Unterstützungspotenziale für den Softwareeinsatz identifiziert. Nach einer Wirtschaftlichkeitsrechnung im individuellen betrieblichen Kontext besteht der nächste Schritt in einer Automation des Prozesses durch einheitliche Anwendung dedizierter, organisatorisch verbindlicher Software, um die Potenziale im Zyklus des Datenqualitätsmanagements zu heben. Revolutionäre Ansätze bestehen in der Anwendung sog. Data-Governance-Software, die extra für diesen Zweck konzipiert und umgesetzt wird.

Eine weitere Herausforderung besteht in den Massendaten (Big Data, z.B. historische oder unstrukturierte externe Daten), die zu sammeln, zu speichern und zu verarbeiten sind. In solchen Unternehmenskontexten ist es aus Effizienz- und Effektivitätsgründen unverzichtbar, entsprechende Software auf dem neuesten Stand der Technik anzuwenden und zusätzliche Kapazitäten aufzubauen.

Begriff	Definition
Eindeutigkeit	Eindeutigkeit steht für den Qualitätszustand, dass eine Instanz ein einziges fachliches Objekt des Unternehmens repräsentiert. Ein Objekt hat eine und nur eine Identität. Eine Kundennummer z.B. repräsentiert einen Kunden „Herr Meier“ und nur einen „Herrn Meier“. Falls die Kundennummer auch noch einen anderen Herrn Meier referenziert, ist der Fakt mehrdeutig.
Genauigkeit	Genauigkeit bezeichnet den Qualitätszustand, dass die Attribute der Daten Werte haben, die im Einklang mit den Merkmalen des Objekts stehen. Die Telefonnummer von Herrn Meier z.B. beinhaltet einen exakten Wert; andernfalls erfüllt sie nicht ihren Zweck und Herr Meier ist nicht telefonisch erreichbar.
Vollständigkeit	Vollständigkeit steht für den Qualitätszustand, dass alle Attribute eines Objekts vorhanden sind und gemessen werden können. Beispielsweise sollten der Name, das Alter und die Adresse jedes Mitarbeiters aus einer Personaldatei ermittelbar sein.
Konformität	Konformität bezeichnet den Qualitätszustand, dass die Attributwerte der Daten im Einklang stehen mit definierten Anforderungen und/oder mit etablierten Standards. Beispielsweise ist der Postleitzahlen-Standard in Deutschland eine fünfstellige Nummer. Jede Abweichung von diesem Standard ist nichtkonform und erfüllt damit nicht die Qualität.
Integrität	Integrität steht für den Qualitätszustand, dass die Beziehungen zwischen Geschäftsobjekten und konkreten Objekten des Unternehmens respektiert werden, z.B. der Datentyp, der Wertebereich oder die Vater-Sohn-Abhängigkeit. Beispielsweise hängt die Prämienverpflichtung eines konkreten Versicherungsvertrags ab von einem Versicherungstarif. Die Versicherungsprämie wird, falls das Versicherungsgeschäft in Deutschland getätigt wurde, als eine Zahl mit zwei Nachkommastellen und dem Währungssymbol € angegeben.
Konsistenz	Konsistenz kann als ein Qualitätszustand definiert werden, bei dem identische Fakten in unterschiedlichen Datenhaltungen auch identisch abgebildet sind. Beispielsweise speichern die Bestandsverwaltungssysteme für die Lebens- und Krankenversicherung beide, dass Herr Meier zwei Kinder hat.

Literatur

CGI (Hrsg.) /BI Framework 2009/
CGI (Hrsg.): The BI Framework. How to Turn Information into a Competitive Asset. O.O. 2009

EIOPA (Hrsg.) /Summary Comments 2009/
EIOPA (Hrsg.): Summary of Comments on CEIOPS-CP-43/09. Consultation Paper on the Draft L2 Advice on TP - Standards for data quality (CEIOPS-SEC-106/09). O.O. 2009
Auf den Seiten der EIOPA
https://eiopa.europa.eu/
gefunden am 12.09.2016

EIOPA (Hrsg.) /Technical Provisions (CP43) 2009/
EIOPA (Hrsg.): CEIOPS' Advice for Level 2 Implementing Measures on Solvency II: Technical Provisions – Article 86 f Standards for Data Quality (former CP 43). Frankfurt, October 2009