Wege aus dem Datenlabyrinth

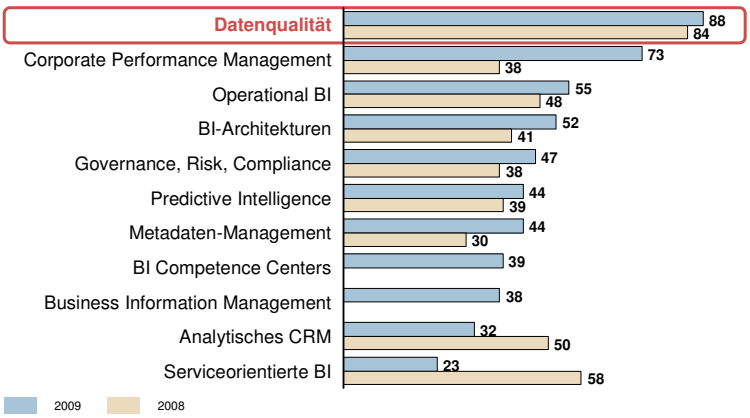
- Datenqualität auf dem Prüfstand -

17. November 2009

Studie IT-Trends 2009 in Deutschland: Das BI-Top-Thema ist Datenqualität

Business Intelligence: Bedeutung einzelner Themen [%]

„Wie wichtig werden die folgenden Themen im Zusammenhang mit Business Intelligence in den nächsten 12 Monaten für Sie sein?“



	Mittelwert	
	2009	2008
Datenqualität	1,9	1,9
Corporate Performance Management	2,2	2,8
Operational BI	2,7	2,6
BI-Architekturen	2,5	3,1
Governance, Risk, Compliance	2,9	3,5
Predictive Intelligence	2,8	2,8
Metadaten-Management	3,2	3,1
BI Competence Centers	3,2	
Business Information Management	3,0	
Analytisches CRM	3,2	2,7
Serviceorientierte BI	4,2	2,4

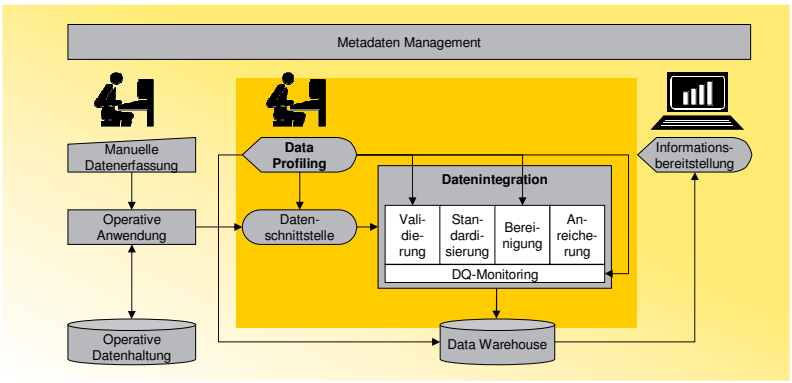
Quelle: Capgemini IT-Trends 2009; Basis: Befragte, die Business Intelligence für eines der 3 wichtigsten Themen halten (n = 35)

tdwi 2 © DATENQUALITÄT ERFOLGREICH STEUERN 2009

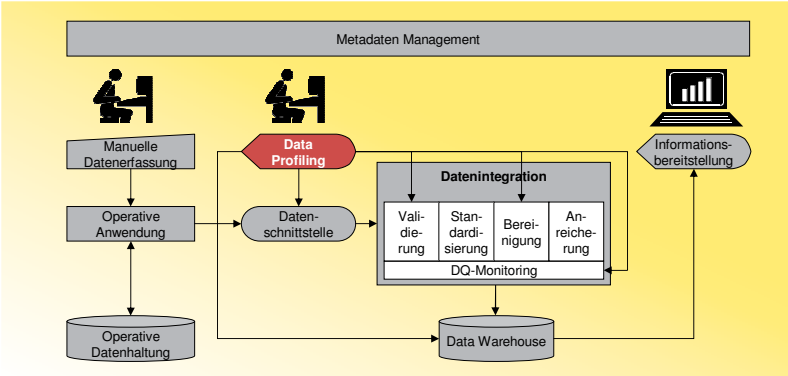
AGENDA

- **Data Profiling**
- **Datenintegrationsprozess**
  - **Validierung**
  - **Standardisierung und Bereinigung von Adressen**
  - **Datenanreicherung**
  - **DQ-Monitoring**
- **Fazit**

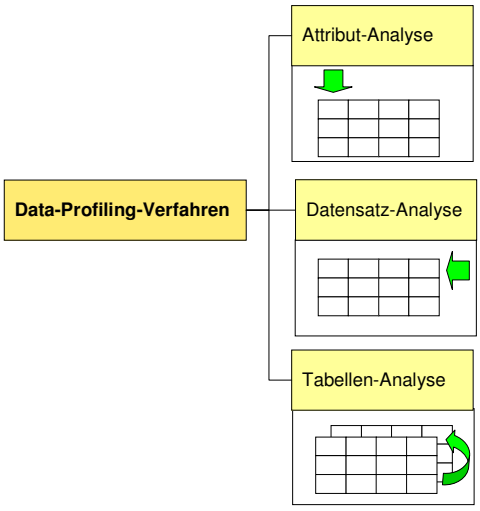
Bausteine eines erfolgreichen Datenqualitätsmanagements



Data Profiling automatisiert den Prozess der Analyse von Dateninhalten



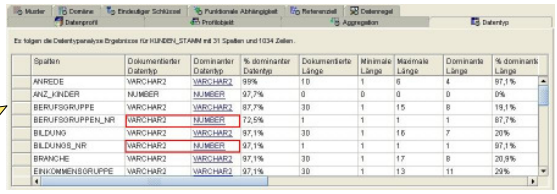
Nur die richtige Kombination der Data-Profiling-Verfahren führt zum Erfolg

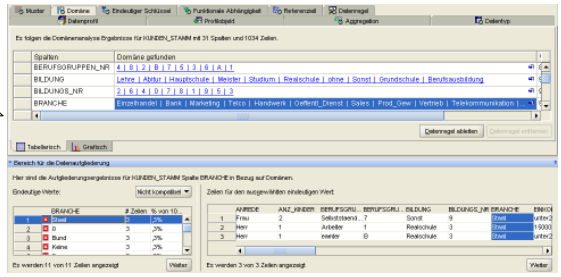


Die Standardverfahren zur Attribut-Analyse entdecken formale Datenfehler

Standardanalysen

- Attributnamen
- Datentyp
- Wertebereich
- Muster
- Domänen
- Eindeutigkeit
- Null-Wert





Mit der Datensatz-Analyse werden funktionale Abhängigkeiten identifiziert

Standardanalysen

- Schlüsselattribute
- Abgeleitete Attribute

Geschäftsregeln

- funktionale Abhängigkeiten

Eindeutiger Schlüssel	Dokumentiert?	Ermittelt?	Lokales Attribut	% Rows
UK_434	Nein	Ja	KUNDENNAME, GEBURTSDATUM	99,5

Nicht eindeutige Datensätze:

KUNDENNAME	GEBURTSDATUM	# Rows	% Rows
Karl Müller	25.09.1966	2	0,2 %
Tabea Schmidt	09.07.1998	3	0,3 %

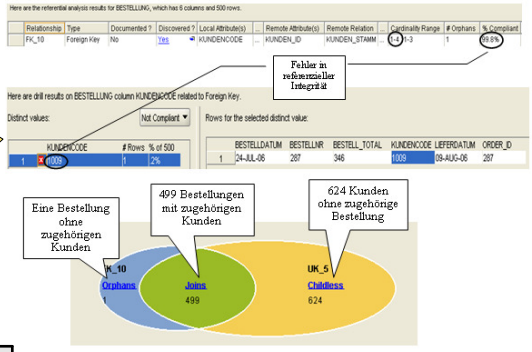
Nicht eindeutige Datensätze zum Kunden „Karl Müller" und Geburtsdatum: „25.09.1966":

KUNDENNAME	GEB_DATUM	ADRESSE	ORT
Karl Müller	25.09.1966	Im Mühlengarten 3	Mechernich
Karl Müller	25.09.1966	Kommerner Straße 8	Euskirchen

Die referenziellen Abhängigkeiten zwischen Tabellen werden zur Fehlersuche genutzt (Tabellen-Analyse)

Standardanalysen

- referenzielle Abhängigkeiten
- Kardinalitäten
- redundante Attribute
- Anzahl Datensätze



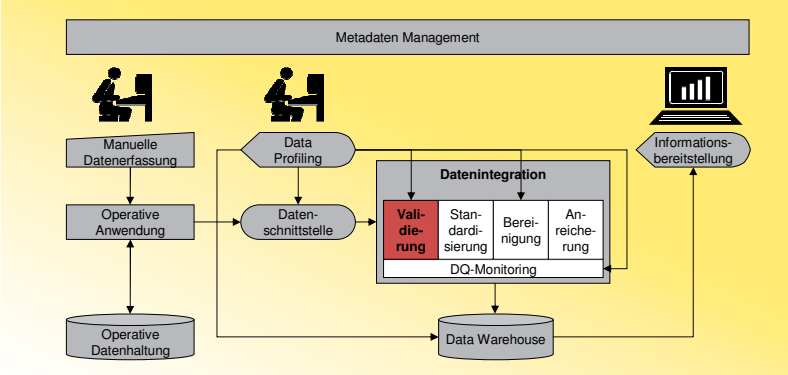
Gruppierung	Wert?	# Zeilen	∅ Zeilen
Monat	Januar 2008	214222	214222
Monat	Februar 2008	237123	225.673
Monat	März 2008	227528	226.291
Monat	April 2008	214789	223.416
Monat	Mai 2008	241927	227.118
Monat	Juni 2008	216456	225.341
Monat	Juli 2008	123546	210.799

Bei der Durchführung des Projekts sollten die vier wichtigsten Faktoren für ein effizientes und erfolgreiches Data Profiling beachtet werden

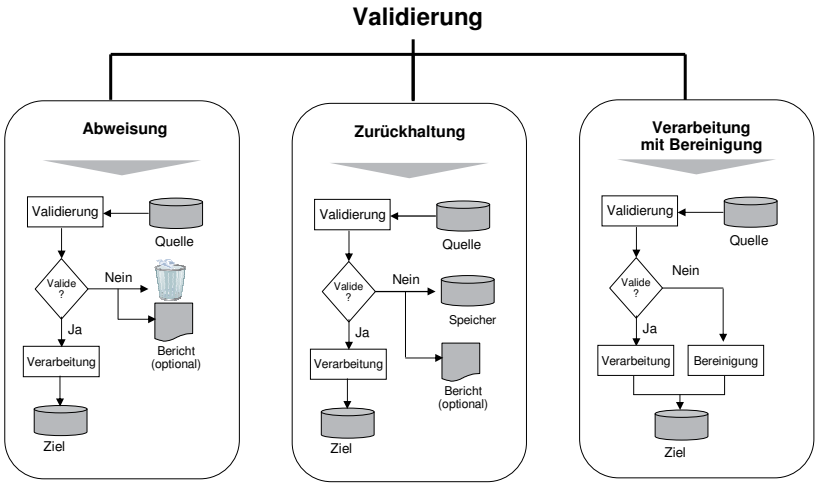
Erfolgsfaktoren

- methodisches Vorgehen
- fundierte Kenntnisse über die Fachlichkeit und die zugehörigen Prozesse
- optimale Zusammensetzung und Erfahrung des Teams
- richtige Kombination und Einsatz der verschiedenen Verfahren

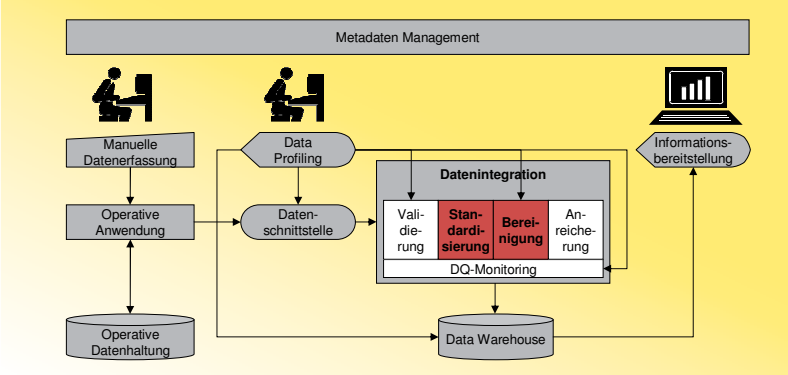
Validierung



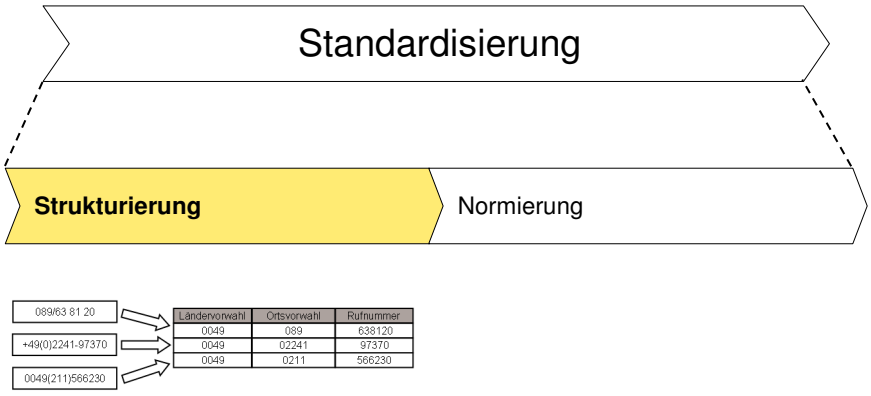
Verfahren zur Validierung



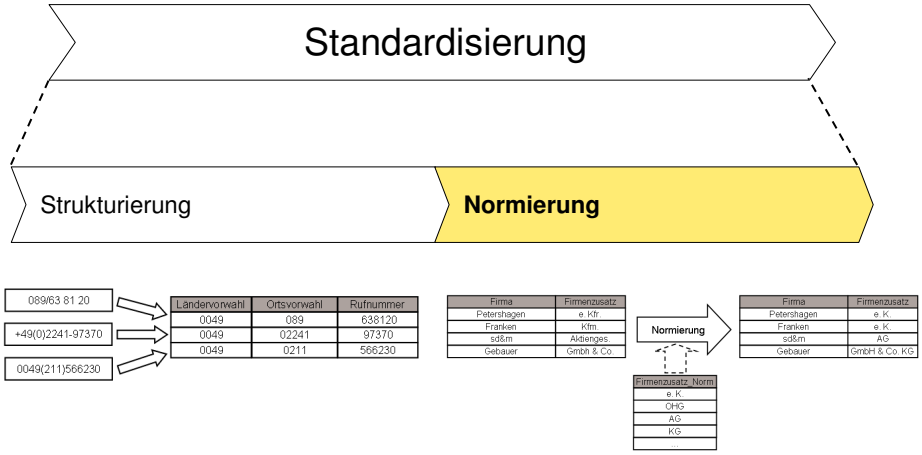
Standardisierung und Bereinigung von Adressen



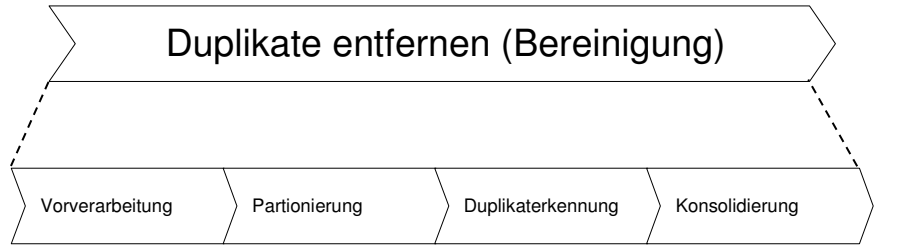
Am besten lassen sich standardisierte Adressen bereinigen



Am besten lassen sich standardisierte Adressen bereinigen



Der Prozess zur Entfernung von Duplikaten gliedert sich in vier Schritte



Die Vorverarbeitung legt den Grundstein für den Erfolg

Vorverarbeitung → Partionierung → Duplikaterkennung → Konsolidierung

- einheitliche Groß-/Kleinschreibung ⇨ Hôtel zum Roß → hôtel zum roß
- Entfernung von Leerzeichen ⇨ hôtelzumroß
- Entfernung von Füllwörtern ⇨ hôtelroß
- Auflösung von Umlauten und & ⇨ hôtelross
- Reduktion diakritischer Zeichen ⇨ hotelross

Die Partionierung steigert die Performance, birgt aber Risiken

Vorverarbeitung → **Partionierung** → Duplikaterkennung → Konsolidierung

Name	Postleitzahl	Ort	Telefon
Meier	53894	Mechernich	02256/99564
Meier	35894	Mechernich	02256/99564
Schmitz	37077	Göttingen	0551/234547
Krüger	53840	Troisdorf	02241/32936

Partitionierungs-schlüssel = 1. Ziffer PLZ

Name	Postleitzahl	Ort	Telefon
Meier	53894	Mechernich	02256/99564
Krüger	53840	Troisdorf	02241/32936

Name	Postleitzahl	Ort	Telefon
Meier	35894	Mechernich	02256/99564
Schmitz	37077	Göttingen	0551/234547

Vergleich | Vergleich

Die Duplikate zu „Meier“ werden durch die Aufteilung in verschiedene Partitionen nicht erkannt!

Zur Erkennung von Duplikaten werden die Adressen miteinander verglichen

Vorverarbeitung | Partionierung | **Duplikaterkennung** | Konsolidierung

Vorname	Name	Strasse	PLZ
Ulrich	Meier	Mülheimer Straße	53840
Ullrich	Mayer	Mülheimer Straße	33480

Vergleichsfunktion

- ✓ „Der Vorname ist an weniger als drei Stellen unterschiedlich„.
- ✓ „Der Nachname ist phonetisch ähnlich"
- ✗ „Die ersten zwei Ziffern der PLZ sind gleich"

Entscheidungsfunktion

Duplikate

Nicht-Duplikate

Vorname	Name	Strasse	PLZ
Ulrich	Meier	Mülheimer Straße	53840
Ullrich	Mayer	Mülheimer Straße	33480

tdwi 19 © DATENQUALITÄT ERFOLGREICH STEUERN 2009

Mit der passenden Metrik lässt sich die Ähnlichkeit messen

Vorverarbeitung | Partionierung | **Duplikaterkennung** | Konsolidierung

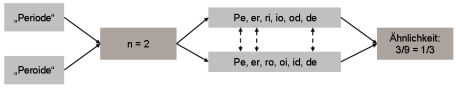
Textuelle Werte

- Phonetik
  - Soundex
  - Kölner Phonetik
- Editierabstand
  - Levenshtein-Distanz
- n-Gramme
- Rekursiver Feldabgleich
- WHIRL
- Ontologien

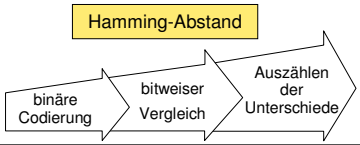
Numerische Werte

- Differenzbildung
- Hamming-Abstand
- Ontologien
- Konvertierung in textuelle Werte

Bi-Gramme (n =2)



Hamming-Abstand



tdwi 20 © DATENQUALITÄT ERFOLGREICH STEUERN 2009

Aus den Duplikaten entsteht ein konsolidierter „Goldener Datensatz“

Vorverarbeitung → Partionierung → Duplikaterkennung → **Konsolidierung**

Konsolidierung

… identischer Duplikate

Maßnahmen

- Löschung der überzähligen Datensätze

Artikel_ID	Bezeichnung	Lieferant_ID
100	Schlüssel, 13 mm	234
102	Schlüssel, 15 mm	243

+

Artikel_ID	Bezeichnung	Lieferant_ID
202	Nuss, VK, 13 mm	649
102	Schlüssel, 15 mm	243

=

Artikel_ID	Bezeichnung	Lieferant_ID
100	Schlüssel, 13 mm	234
102	Schlüssel, 15 mm	243
202	Nuss, VK, 13 mm	649

… ergänzender Duplikate

- Zusammenführung der Datensätze

Artikel_ID	Grp_ID	Bezeichnung	Preis_Netto
100	1	Schlüssel, 13 mm	4,39
102	1	Schlüssel, 15 mm	4,73
202	1	Nuss, VK, 13 mm	3,67

+

Artikel_ID	Bezeichnung	Lieferant_ID
100	Schlüssel, 13 mm	234
102	Schlüssel, 15 mm	243
202	Nuss, VK, 13 mm	649

=

Artikel_ID	Grp_ID	Bezeichnung	Preis_Netto	Lieferant_ID
100	1	Schlüssel, 13 mm	4,39	234
102	1	Schlüssel, 15 mm	4,73	243
202	1	Nuss, VK, 13 mm	3,67	649

…widersprechender Duplikate

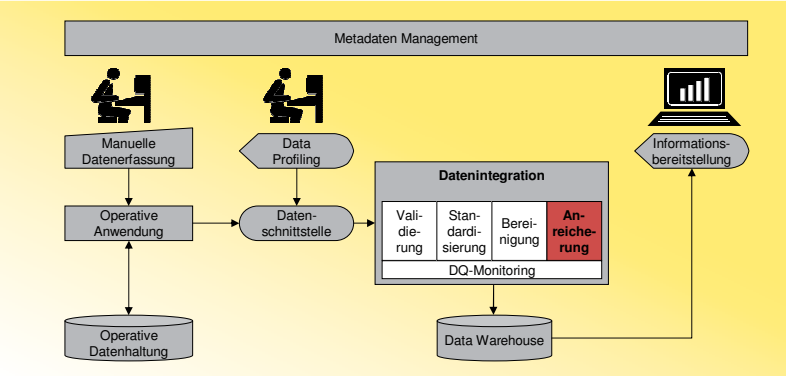
- Konfliktvermeidende Verfahren
- Konfliktlösende Verfahren

Artikel_ID	Grp_ID	Bezeichnung	Preis_Netto
100	1	Schlüssel, 13 mm	4,39
102	1	Schlüssel, 15 mm	4,73
202	1	Nuss, VK, 13 mm	3,67

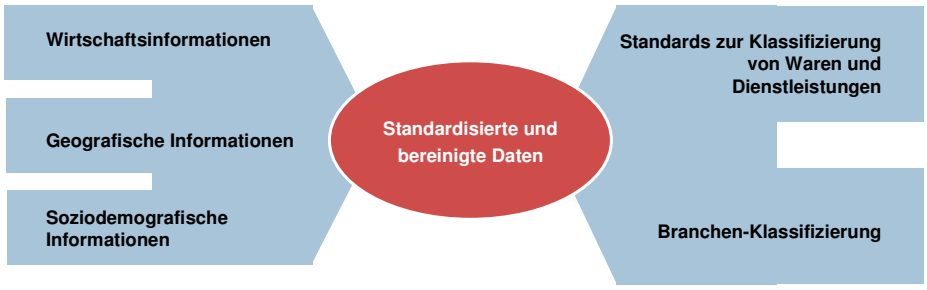
+

Artikel_ID	Bezeichnung	Preis_Netto
100	Schlüssel, 13 mm	3,89
102	Schlüssel, 15 mm	4,02
202	Nuss, VK, 13 mm	3,56

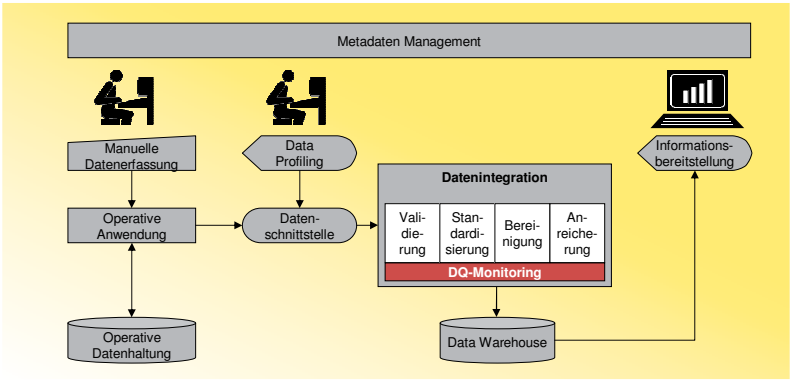
Datenanreicherung



Daten werden angereichert um ihren Informationswert zu steigern

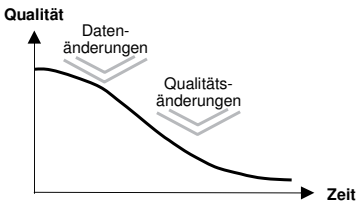


Fortlaufendes Monitoring der Datenqualität sichert den nachhaltigen Erfolg des Datenqualitätsmanagements

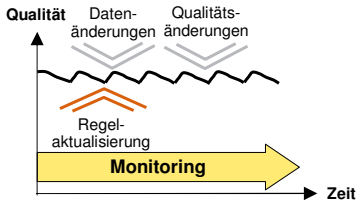


Voraussetzung für die Erhaltung der erreichten Datenqualität ist fortlaufendes Monitoring

Datenqualitäts-Lifecycle (ohne Monitoring)



Datenqualitäts-Lifecycle (mit Monitoring)



Das Monitoring im Betrieb muss automatisiert erfolgen.

DQ-Kennzahlen aus dem Monitoring werden den DQ-Kriterien zugeordnet

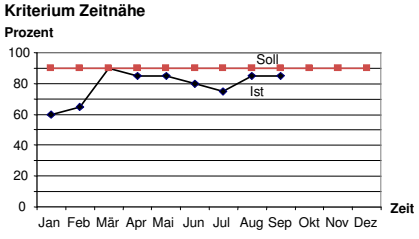
Darstellungsformen der Ergebnisse aus dem Monitoring

Ampeldarstellung

	SOLL	IST	AMPEL
Korrektheit	95%	96%	○ ○ ●
Vollständigkeit	90%	88%	○ ● ○
Zuverlässigkeit	95%	75%	● ○ ○

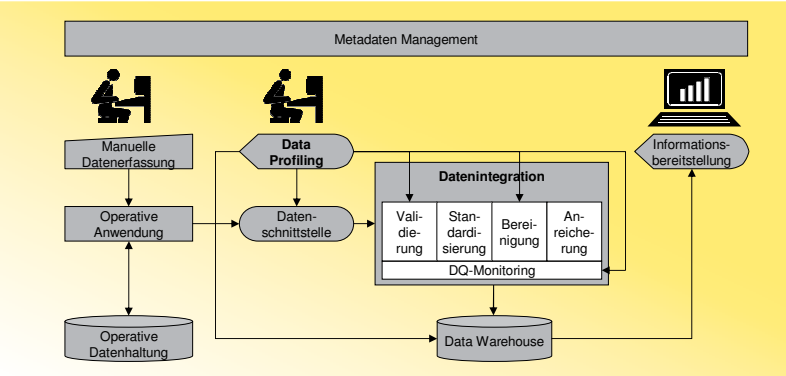
DQ-Kennzahlen/-Kriterien werden in Bezug auf Schwellwerte betrachtet analog zu Dashboards im Umfeld von Business Intelligence.

Zeitreihenbetrachtung



Die zeitliche Entwicklung von DQ-Kennzahlen/-Kriterien wird betrachtet, häufig mit Angabe einer Ziellinie oder eines Toleranzbereichs.

Bausteine eines erfolgreichen Datenqualitätsmanagements



Produktauswahl

Kategorie / Anbieter	Data Profiling	Daten-validierung	Daten-bereinigung	Name & Adresse	Dubletten	DQ-Monitoring
Oracle	X	X	(OEM Trillium)	X	X	
IBM	X	X	X	X	X	X
Informatica	X	X	X	X	X	X
SAP	X	X	X	X	X	X
DataFlux	X	X	X	X	X	X
Human Inference	X	X	X	X	X	X
Trillium	X	X	X	X	X	X
Uniserv	X	X	X	X	X	X
Innovative Systems	X	X	X	X	X	
Pitney Bowes (Group 1)	X	X	X	X	X	X

Fazit

Die Hoffnung vieler Unternehmen auf Lösung dieses Problems durch die bloße Einführung neuer Systeme oder standhaftem Ignorieren der Problematik schwindet und macht endlich Platz für **wirksame Maßnahmen**.

- Der Schlüssel zum Erfolg ist ein proaktives **Datenqualitätsmanagement**.

- Insbesondere das **Data Profiling** erkennt viele Schwachstellen in den Quelldaten und sollte zu Beginn eines Projektes eingesetzt werden.
- Durch **Standardisierung, Bereinigung** und **Anreicherung** der Daten kann deren Wert für ein Unternehmen sehr stark erhöht werden.
- Das **Data Quality Monitoring** bietet im produktiven Betrieb die Sicherheit, dass Veränderungen und neue Qualitätsprobleme in den Daten rechtzeitig erkannt werden.

Es gibt keine Ausreden mehr für schlechte Daten – fangen Sie an!

Die Expertise zum Datenqualitätsmanagement in BI-Projekten wurde in einem Fachbuch gebündelt

Detlef Apel, Wolfgang Behme, Rüdiger Eberlein, Christian Merighi
"Datenqualität erfolgreich steuern"

Buchteil:	Theorie	Technische Umsetzung	Projektpraxis
Inhalte:	• Datenqualität • Ursachen und Ausprägungen schlechter Datenqualität • Auswirkungen schlechter Datenqualität • Organisation • Referenzarchitektur für Business-Intelligence-Anwendungen • Kennzahlen zur Messung der Datenqualität	• Verbesserung der Datenqualität im Quellsystem • Data Profiling • Erfolgreiche Datenvalidierung und -filterung • Standardisierung und Bereinigung • Datenanreicherung • Verbesserung der Datenqualität in der Bereitstellung und Präsentation • Metadaten Management • Data Quality Monitoring • Produktauswahl/-integration	• Datenqualitätsmanagement in einer Studie • Datenqualitätsmanagement in der Spezifikation • Datenqualitätsmaßnahmen in der Konstruktionsphase • Steuerung der Datenqualität in der Realisierung • Steuerung der Datenqualität im Betrieb