

# DATENQUALITÄT IN DATA-WAREHOUSE SYSTEMEN -TEIL 1

## DAS ENDE VON „GARBAGE IN - GARBAGE OUT“

Schon längst haben sich Data Warehouse Systeme als systemseitige Grundlage für Business Intelligence Strategien etabliert und werden von vielen Unternehmen eingesetzt, um eine anwender- sowie zeitnahe Entscheidungsunterstützung zu gewährleisten. Um aus dem hier gewonnenen Wissen korrekte operative und strategische Entscheidungen ableiten zu können, ist eine hohe Qualität der Datenbasis Grundvoraussetzung. Nur wenn die analysierten Daten bereinigt und frei von Mängeln sind, können damit qualitativ hochwertige Entscheidungen getroffen werden, die zu einem echten Mehrwert für das Unternehmen führen.

Oftmals weisen die für BI benötigten Daten erhebliche Qualitätsdefizite auf, welche sich auf die Heterogenität der Quellsysteme, menschliche Schwächen oder auch die beteiligten Prozesse zurückführen lassen. Der Grund sind fehlende oder falsche Werte, unterschiedliche Wertausprägungen für identische Objekte (Synonyme/Homonyme) oder auch doppelte und veraltete Daten.

Jeder kennt das Problem: Heinrich Mayer, Heinrich Maier (leider falsch), Heinrich K. Mayer, H. Mayer, Heinrich Kurt Mayer. Alles ein und dieselbe Person. Aber weiß das auch die Datenbank?

Um zu verhindern, dass die zentrale Datenhaltung im Data Warehouse System durch qualitativ mangelhafte Daten kontaminiert wird, muss das Thema Datenqualität als Bestandteil der Architektur eines DWHs verstanden und

in der Organisation eines Unternehmens verankert werden.

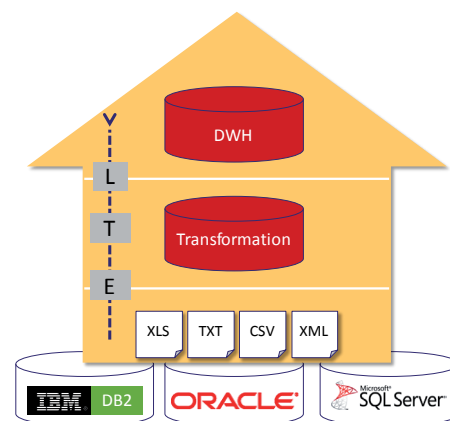
Als Basis für diese Aktivitäten stellen Data Warehouse Systeme in diesem Kontext eine zentrale Komponente dar. Durch sie werden heterogene Datenbestände aus unterschiedlichen,

sowohl unternehmensinternen als auch externen IT-Systemen, konsolidiert und ermöglichen somit systemübergreifende Auswertungen und Analysen.

### DATENQUALITÄT - DEFINITION

Ähnlich wie in der industriellen Fertigung, in der die Qualität eines Gegenstandes oder Produktes durch Messung der Eigenschaften anhand von Sollwerten und Toleranzbereichen überprüft wird, so empfiehlt es sich auch bei Daten eine Messung bzw. Evaluierung der Eigenschaften vorzunehmen.

Daten und Informationen eines Unternehmens müssen in der Regel für unterschiedliche Anwendungs- und Nutzungsszenarien zur Verfügung stehen. Daher ist es notwendig, anwendungsabhängige Qualitätsmerkmale zu definieren, die durch geeignete Metriken beurteilt werden können.



Demnach lässt sich der Begriff der Datenqualität definieren als Grad, in dem ein Satz Eigenschaften / Qualitätsmerkmale eines Datenprodukts konkrete Anforderungen erfüllt.

Im Data Warehouse Kontext bedeutet dies, dass die zu integrierenden Daten der Quellsysteme in allen Phasen des Data Warehousing anhand von festgelegten Qualitätsmerkmalen mittels entsprechender Metriken zu überprüfen bzw. zu beurteilen sind. Nicht die Technik bzw. das Tool macht die Datenqualität, sondern das methodische Vorgehen.

### DATA PROFILING

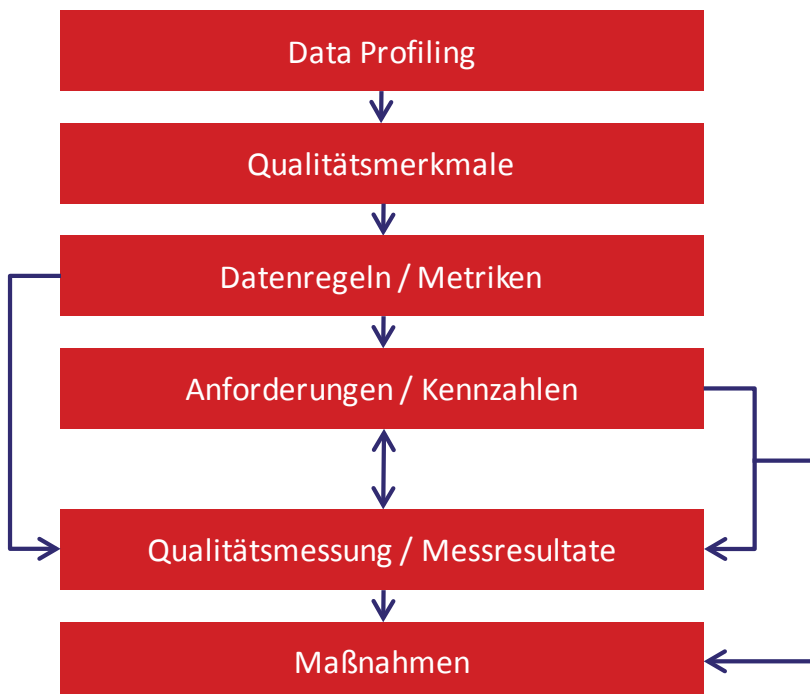
Damit Datenqualität gemessen bzw. beurteilt werden kann, sind im Vorfeld die zu überprüfenden Merkmale und Eigenschaften der Daten festzulegen. Durch das sogenannte Data Profiling werden BI-relevante Datenbestände mittels unterschiedlicher Analysen untersucht, und deren Qualitätsmerkmale ermittelt.

Der Zweck bzw. die Idee dahinter ist, über den tatsächlichen Inhalt der zu analysierenden Daten die eigentliche Struktur zu verstehen und somit eine Aussage

über die Qualität der künftigen Datenlieferungen treffen zu können. Dabei werden die dokumentierten Metadaten der relevanten Datenbestände mit den eigentlichen Produktivdaten validiert und darüber hinaus neue Metadaten identifiziert. Metadaten im Allgemeinen sind Daten, die ein Betrachtungsobjekt beschreiben. Im Kontext der Datenqualität beschreiben sie die Struktur sowie die technischen und betriebswirtschaftlichen Zusammenhänge.

Data Profiling spiegelt somit einen Prozess wieder, in dem es darum geht, Daten und deren Strukturen zu verstehen. Somit werden keine Datenqualitätsprobleme behoben, sondern vielmehr die Grundlage für ein Regelwerk geschaffen, das Datenqualitätsprobleme erkennen kann.

Die vom PENTASYS Team eingesetzten ETL-Werkzeuge bieten umfassende Möglichkeiten sogenannte Datenprofile zu erstellen. Dabei werden eine Vielzahl von Analysemöglichkeiten auf Attributs-, Datensatz- sowie Tabellenebene genutzt, um letztendlich ein genaues Bild über die qualitätsrelevanten Merkmale und Eigenschaften der zu integrierenden Datenbestände erzeugen zu können.



### Data Profiling im Data Warehouse Kontext

Data Profiling liefert wichtige Aspekte für die Bewertung der Datenqualität. Im Data Warehouse Kontext ist Data Profiling zielgerichtet in den jeweiligen Phasen bzw. Lebenszyklen einzusetzen.

Prinzipiell kann zwischen dem initialen und dem laufenden Data Profiling unterschieden werden. Das initiale Data Profiling wird intensiv bei der Realisierung eines Data Warehouse Systems betrieben, um die Beschaffenheit der zu integrierenden Datenbestände beurteilen zu können. Das laufende Data Profiling wird regelmäßig zur Kontrolle der Datenqualität eingesetzt. Grundsätzlich gilt es, sämtliche Ergebnisse die im Data Profiling ermittelt worden sind, fachlich zu überprüfen.

## MESSUNG DER QUALITÄT MIT DATENREGELN UND KENNZAHLEN

Um Datenqualität letztendlich messen zu können ist es erforderlich, überprüfbare Anforderungen zu definieren. Die Grundlage für entsprechende Anforderungen liefern die im Data Profiling ermittelten Qualitätsmerkmale der Daten.

Über diese Merkmale müssen die Anforderungen der Anwender identifiziert und in überprüfbare Vorgaben transformiert werden. Dies geschieht durch die Definition von formalen sowie fachlichen Datenregeln mit entsprechenden Kennzahlen. Formale Datenregeln beschreiben eher die Struktur der Daten, fachliche Datenregeln hingegen überprüfen die betriebswirtschaftlichen Zusammenhänge.

Durch sie wird es erst möglich eine Metrik festzulegen, die Qualitätskennzahlen als Ergebnis liefert und somit die Beschaffenheit der Daten beschreibt. So könnte beispielsweise eine formale Datenregel lauten, dass für ein Attribut Geb\_Datum der Relation Kunde das Werte-Muster in der Form DD.MM.JJJJ vorliegen muss.

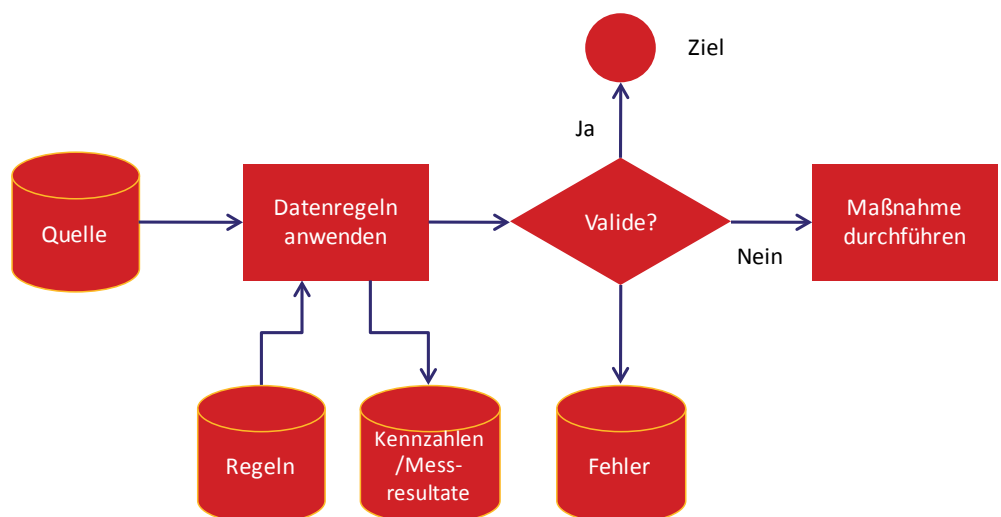
Die daraus resultierende Metrik ist das Verhältnis der Anzahl der Datensätze welche ein unzulässiges Werte-Muster vorweisen, zur Gesamtanzahl der überprüften Datensätze. Die Kennzahl wäre entsprechend ein Verhältniswert, welcher in Prozent dargestellt werden kann und die Anforderungen der Anwender widerspiegelt.

Die Definition von Datenregeln ist die wichtigste und zugleich zeitaufwändigste Aktivität im Datenqualitätsmanagement. Datenregeln können aus den Ergebnissen des Data Profiling abgeleitet bzw. aus den existierenden Geschäftsregeln übernommen werden. Da Geschäftsregeln oftmals nicht ausreichend bzw. zum Teil gar nicht dokumentiert sind, ist es wichtig, dass der jeweilige Fachbereich sein Wissen bei der Regelbildung einbringt. Die gängigste Art Regeln umzusetzen ist das Entwickeln von entsprechenden SQL-Skripts.

Die gängigen Datenbanken bieten in diesem Kontext individuelle Lösungsmöglichkeiten, einfache bis sehr komplexe Datenregeln zu entwickeln und anzuwenden. Diese Regeln werden meist als Datenbankobjekte innerhalb der Datenbank gespeichert und verwaltet.

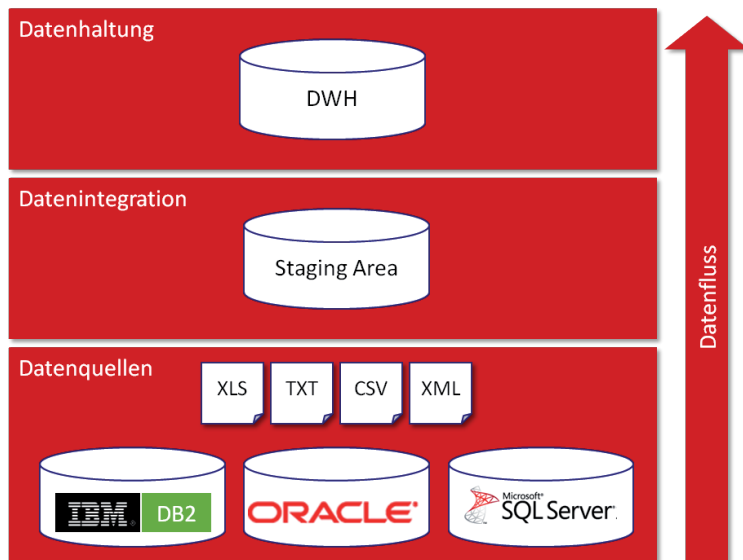
Sobald ein Regelwerk existiert kann mit der Qualitätsmessung begonnen werden. Bei der Durchführung der Qualitätsmessung werden die spezifizierten Regeln auf den betrachteten Datenbestand angewendet. Existiert eine Regelverletzung, so sind die fehlerhaften Daten in einem extra dafür vorgesehen Speicherbereich auszusortieren. Hier können diese dann nochmals analysiert und gegebenenfalls korrigiert werden.

Die ermittelten Messresultate sind abschließend an den zuvor definierten Qualitätskennzahlen zu spiegeln, um die eigentliche Beschaffenheit der Datenqualität erkennen zu können.



## DATENQUALITÄT – MESSBEREICHE INNERHALB DER REFERENZARCHITEKTUR

Nachdem der Begriff der „Datenqualität“ geklärt ist und aufgezeigt wurde, wie Datenqualität gemessen werden kann, werden im zweiten Teil dieses Artikels mögliche Messbereiche innerhalb einer BI-Referenzarchitektur behandelt. Die folgende Abbildung zeigt sehr abstrakt die grundlegenden Schichten einer möglichen Referenzarchitektur.



Im zweiten Teil wird dann speziell auf diese Messbereiche und die Datenbereinigung eingegangen.

### Über den Autor

Christian Brunner ist als Consultant in den Bereichen Business Intelligence sowie dem Qualitätsmanagement in Softwareprojekten bei der Pentasys AG tätig. Sein fachlicher und technischer Fokus liegt dabei auf der Konzeption und Entwicklung von Data Warehouse Systemen sowie Analyse- und Reporting-Applikationen.



blickpunkte – Das Magazin rund um IT-Themen ist ein kostenloser Newsletter der PENTASYS AG



## PENTASYS

Unser Maßstab ist der Mensch

Die PENTASYS AG gehört zu den am schnellsten wachsenden deutschen IT-Systemintegratoren. 1995 mit drei Beschäftigten gegründet, hat das Unternehmen bis heute mehr als 200 neue qualifizierte Arbeitsplätze in Deutschland geschaffen. Die Unternehmensstrategie ist auf kompromisslose Qualität und strikte Orientierung am Mehrwert für die Kunden ausgerichtet. Hochqualifizierte und überdurchschnittlich motivierte Mitarbeiter sowie ein gemäß ISO-9001/2008 zertifiziertes Projektvorgehensmodell schaffen die Voraussetzungen hierfür.

Zum Leistungsspektrum gehören Consulting, Projektmanagement, Machbarkeitsanalyse, Architekturkonzeption, Realisierung und Test von IT-Systemen aus einer Hand. Zu den Referenzkunden zählen unter anderem ADAC e.V., Arval (eine BNP Paribas Company), CACEIS Bank, Deutsche Bahn AG, DekaBank Deutsche Girozentrale, Deutsche Post AG, Deutsche Telekom, BMW AG, Direkt Anlage Bank, Bristol-Myers Squibb, MAN Truck & Bus AG, Telefonica Germany GmbH & Co. OHG, RTL II, TÜV Süd AG, Yves Rocher, Volvo Financial Services, das ifo-Institut für Wirtschaftsforschung und das Europäische Patentamt.

#### COPYRIGHT:

Alle Inhalte, auch Konzepte und Design, des Newsletters sind urheberrechtlich geschützt. Das Copyright/Urheberrecht liegt dabei bei der PENTASYS AG.

Das Zitieren ist unter Berücksichtigung der üblichen Regeln und Hinweise gestattet. Das Kopieren oder der Nachdruck, auch auszugsweise, sowie fotomechanische Wiedergabe oder Erfassung auf Datenträgern ist nur mit schriftlicher Genehmigung der PENTASYS AG zulässig.

Sofern in den vorliegenden Inhalten Marken und geschäftliche Beziehungen verwendet werden, auch wenn diese nicht als solche gekennzeichnet sind, gelten die entsprechenden Schutzbestimmungen.

#### KONTAKT:

##### PENTASYS AG

Rüdesheimer Straße 9  
80686 München  
Tel.: (0 89) 5 79 52-0  
Fax: (0 89) 5 79 52-399

##### PENTASYS AG

Geschäftsstelle Frankfurt  
Solmsstraße 41  
60486 Frankfurt am Main  
Tel.: (0 69) 7 07 98 39-0  
Fax: (0 69) 7 07 98 39-5 99

##### PENTASYS AG

Geschäftsstelle Köln  
Dülkenstraße 9  
51143 Köln  
Tel.: (0 22 03) 9 35 48 -76  
Fax: (0 22 03) 9 35 48 -78

redaktion@pentasys.de  
www.pentasys.de