

DATENQUALITÄT IN DATA-WAREHOUSE SYSTEMEN - TEIL 2

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 auch qualitativ hochwertige Entscheidungen getroffen werden, die zu einem echten Mehrwert für ein Unternehmen führen.

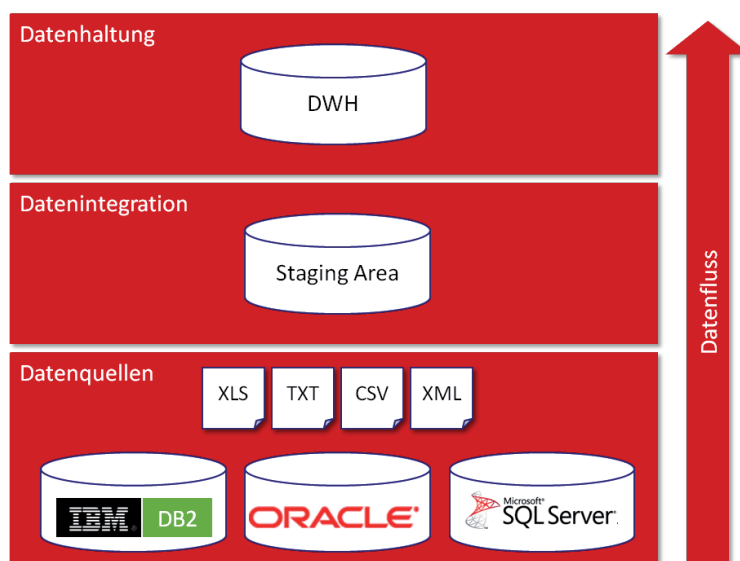
Nachdem im Teil 1 dieses Artikels der Begriff der „Datenqualität“ geklärt und aufgezeigt wurde, wie Datenqualität gemessen werden kann, wird nun im Folgenden auf mögliche Messbereiche innerhalb einer BI-Referenzarchitektur eingegangen. Die nebenstehende Abbildung abstrahiert die grundlegenden Schichten einer möglichen Referenzarchitektur.

Wie die Abbildung zeigt, können die drei wesentlichen Bereiche Datenquellen, Datenintegration sowie Datenhaltung

identifiziert werden. Betrachtet man ein Data Warehouse System unter einem

datenflussorientierten Blickwinkel, so stellen die Datenquellen den Ursprung des Datenflusses dar, welcher im Verlauf den Datenintegrationsbereich als Station beinhaltet und letztendlich im Datenhaltungsbereich endet. Der Auslöser für diesen Datenfluss ist der sogenannte ETL-Prozess. Hierbei sind sämtliche Komponenten der Referenzarchitektur beteiligt, die im direkten Zusammenhang mit der Integration der relevanten Quelldaten stehen. Die Reihenfolge der einzelnen Prozessschritte bestimmt dabei den Namen dieses Prozesses.

Zunächst werden die relevanten Daten aus den Quellen in die Staging Area des Datenintegrationsbereichs



extrahiert (E), was einem Transport der Daten entspricht. Anschließend erfolgt in der Staging Area die Transformation (T) der Daten. Unter Transformation versteht man hier sämtliche Aktivitäten, die notwendig sind, um die Daten in ein für das Data Warehouse System geeignetes Format zu überführen sowie die Datenqualität an die neuen Benutzerbedürfnisse anzupassen.

Nachdem die Daten in ein einheitliches Format überführt und bereinigt worden sind, erfolgt anschließend das Laden (L) der Daten vom Integrationsbereich in den Datenhaltungsbereich. Anhand dieses Datenflusses wird im Folgenden auf mögliche Messbereiche für Datenqualität eingegangen:

Messbereich 1 – Datenquellen

Die Implementierung von Prüf- und Messmechanismen direkt in den Quellsystemen hängt wesentlich von deren Flexibilität und technischen Machbarkeit ab. Neben den Fragen nach den technischen Implementierungsmöglichkeiten spielen auch die Auswirkungen auf Performance und Kosten eine zentrale Rolle. Fest steht jedoch, dass Qualitätsprüfungen so früh wie möglich im Datenfluss den größten Nutzen erzielen. Besonders sinnvoll in diesem Bereich sind Prüfungen, welche die Erzeugung bzw. Erfassung der Daten betreffen.

Messbereich 2 – Extraktion / Datenintegrationsbereich

Die Staging Area als zentrale Datenbank des Datenintegrationsbereichs ist die erste Station innerhalb des Data Warehouse Systems, in der die Quelldaten extrahiert und für die Transformation zwischengespeichert werden. Bevor die eigentliche Transformation der Daten erfolgt, können hier sehr gut formale Datenregeln zum Einsatz kommen, welche beispielsweise die Überprüfung der Formate und Datentypen sowie einfache Integritätsvalidierungen beinhalten.

Messbereich 3 – Laden / Datenhaltungsbereich

Nachdem die Daten in der Staging Area transformiert worden sind, erfolgt das Laden der Daten in den Datenhaltungsbereich. Je nach Architekturvariante können hier unterschiedliche Datenhaltungskomponenten zum Einsatz kommen. In der Regel existiert jedoch ein sogenanntes Core Data Warehouse, welches sämtliche BI-relevanten Daten aus allen eingebundenen Unternehmensbereichen zeitübergreifend durch die sogenannte Historienbildung vorhält.

Da die Daten durch die Transformation nun standardisiert vorliegen und somit sehr gut vergleichbar sind, können zum Einen Duplikate besser erkannt und beseitigt werden, zum Anderen können komplexere, fachliche Datenregeln zum Einsatz kommen, welche die betriebswirtschaftlichen Zusammenhänge der Daten system- sowie zeitübergreifend überprüfen.

DATENBEREINIGUNG

Werden im Rahmen der Messung der Datenqualität fehlerhafte Daten identifiziert, müssen diese entsprechend korrigiert werden, was zur Phase der Datenbereinigung führt (engl. Data Cleaning oder Data Cleansing). Die Datenbereinigung ist ein Prozess, der die nachträgliche Wiederherstellung einer korrekten Datenbasis zum Ziel hat. Grundsätzlich kann zwischen der manuellen und der automatisierten Datenbereinigung unterschieden werden, was zu folgenden Fehlerklassen bzgl. Datenqualitätsmängel führt:

1. Klasse – automatische Erkennung und automatische Korrektur

Bei Mängeln dieser Klasse handelt es sich um vorhersehbare, bereits bekannte Fehler, welche sich nach spezifizierten Regeln korrigieren lassen. Ähnlich wie bei der Messung der Datenqualität kommen hier spezielle Bereinigungsroutinen in Form von SQL-Skripts zum Einsatz, welche als Bestandteil des ETL-Prozesses gesehen werden können.

2. Klasse – automatische Erkennung und manuelle Korrektur

Fehler dieser Klasse können zwar automatisiert durch entsprechende Datenregeln erkannt und die betroffenen Datensätze aussortiert werden, erfordern jedoch eine manuelle Korrektur, welche durch den jeweiligen Fachbereich zu erfolgen hat.

3. Klasse – manuelle Erkennung und manuelle Korrektur

In der Praxis ist eine absolut fehlerfreie Datenbasis in der Regel nicht möglich, sodass ein gewisser Rest an Fehlern bestehen bleibt. Diese können nur durch die Anwender selbst manuell erkannt und entsprechend manuell korrigiert werden.

Prinzipiell gilt es bei Bereinigungsmaßnahmen die entstehenden Kosten dem möglichen Nutzen gegenüberzustellen. In vielen Fällen kann eine symptomorientierte Bereinigung ausreichend und wirtschaftlich sinnvoll sein. Sind jedoch die Kosten der Datenbereinigung höher als der dabei erzeugte Nutzen, so sollte über eine Beseitigung der Fehlerursache nachgedacht werden. Eine Beseitigung der Fehlerursache kann kurzfristig zu hohen Kosten führen, diese jedoch amortisieren sich im Laufe der Zeit über die ausbleibende symptomorientierte Datenbereinigung.

DATENQUALITÄTSMANAGEMENT

Nur durch eine sinnvolle Kombination der hier vorgestellten technischen Verfahren mit einer wohldefinierten Organisationsstruktur können Unternehmen nachhaltig ein hohes Datenqualitätsniveau erreichen.

Das Datenqualitätsmanagement befasst sich in diesem Sinne mit der Erfassung, Speicherung und Verwaltung von qualitativ hochwertigen Daten und beinhaltet dabei die notwendigen Rollen und Prozesse, damit das Thema Datenqualität entsprechend in einer Organisation verankert werden kann. Im Folgenden soll daher abschließend auf die wesentlichen Prozesse des Datenqualitätsmanagement kurz eingegangen werden.



Qualitätsplanung

Hier werden die Bedürfnisse und Erwartungen der Anwender erfasst und in konkrete Vorgaben transformiert. Dies kann beispielsweise die Bildung von formalen und fachlichen Datenregeln und deren Qualitätskennzahlen beinhalten, was einer Transformation der Anforderungen entspricht.

Qualitätslenkung

Die Qualitätslenkung steht für die Überwachung der Einhaltung von Vorgaben, was bedeutet, dass die zuvor festgelegten Qualitätsregeln kontinuierlich auf die zu integrierenden Datenbestände angewendet und die Ergebnisse der Qualitätsmessung an den festgelegten Qualitätskennzahlen gespiegelt werden.

Qualitätssicherung

Der Prozess der Qualitätssicherung kann als unterstützende Komponente für die Datenqualitätsplanung und -lenkung verstanden werden. Das bedeutet, dass zuvor erkannte Datenqualitätsprobleme auf deren Ursachen und Auswirkungen hin untersucht werden, was einer Problem- und Risikoanalyse entspricht.

Qualitätsverbesserung

Bei der Qualitätsverbesserung werden konkrete Maßnahmen eingeleitet, die auf den Ergebnissen der Problem- und Risikoanalyse basieren und eine reaktive oder präventive Qualitätsverbesserung zum Ziel haben. Dabei können sowohl Prozessoptimierungen, einmalige Korrekturen und Bereinigungen als auch die Anpassung der Prüflöge Gegenstand der Maßnahme sein.

FAZIT

Daten, die für einen bestimmten operativen Verwendungszweck modelliert worden sind und eine hohe Heterogenität vorweisen, müssen standardisiert und bereinigt werden, um als Basis für entsprechende Analysen im BI-Kontext verwendet werden zu können. Grundsätzlich sollte eine Qualitätssicherung so früh wie möglich im Datenfluss ansetzen.

Dies ist jedoch oft nur eingeschränkt möglich. Daher ist es umso wichtiger, dass eine Sicherstellung der Datenqualität während der Durchführung des ETL-Prozesses vollzogen wird. Die hier vorgestellten technischen Verfahren und Vorgehensweisen sollten dabei mittels entsprechender Prozesse und Rollen durch das Datenqualitätsmanagement in die Organisation eines Unternehmens eingebettet werden.

Das Sicherstellen der Datenqualität ist somit der erste Schritt auf dem Weg zum erfolgreichen Master Data Management.

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, Telefónica o2 Germany, 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