

JETZT MIT **itfokus**

itmanagement



itmanagement

APRIL 2010

10 EURO

www.it-daily.net

**Konferenz-
programm
im Heft**



VORSICHT FALLE!

Wohin im
Lizenzdschungel?

CLOUD COMPUTING

Sicherheit im Fokus

AKTIVE STRATEGIEN

Corporate Data
Quality Management

INTELLIGENTE KENNZAHLEN

Fit für den Wettbewerb

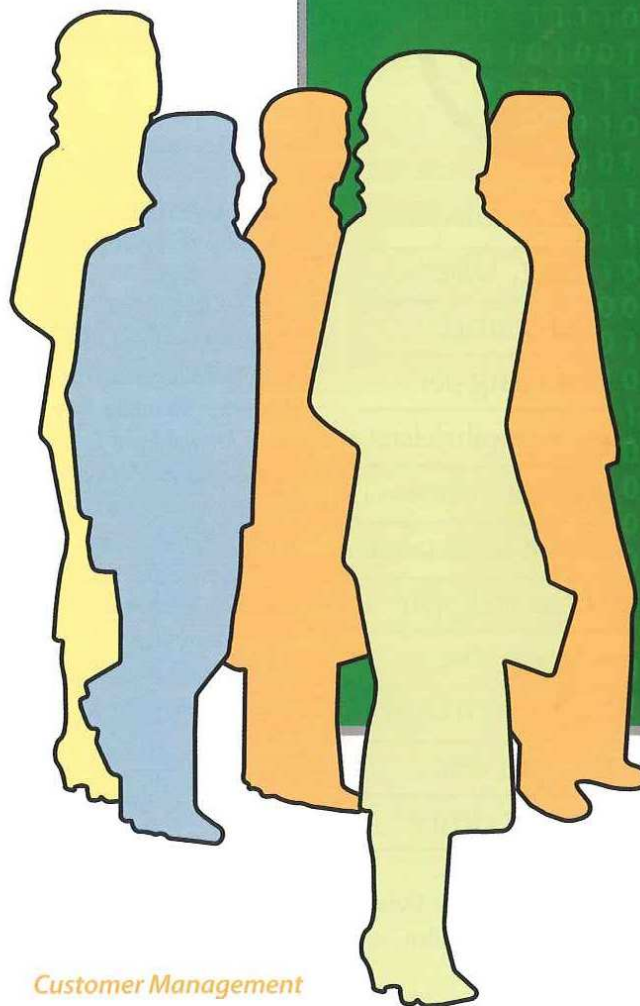
BUSINESS PROCESS EXCELLENCE

DR. PETER KÜRPICK
über die Roadmap der
Software AG und den
Erfolg am Markt

itverlag

▶▶ CARVE OUT-PROJEKTE IM SAP-BEREICH ◀◀

Schlechte Datenqualität kostet Geld, Kunden sowie Vertrauen nach außen und nach innen. Im Customer Management wird dies mit dem Beispiel der schlechten Adressqualität wieder und wieder vor Augen geführt. Weniger leicht zu entdecken, subtiler und nicht weniger kostspielig sind falsche Kundenscores und suboptimale Zielgruppenselektionen, die auf fehlerhaften Data Warehouse-Daten gerechnet und selektiert wurden.



Customer Management

Data Quality Management –

Das Umsatzpotenzial einer Kampagne wird nicht ausgeschöpft, teure Kampagnenressourcen werden fehlgeleitet und im schlechtesten Falle werden Kunden mit unpassenden Angeboten vergrault. Data Warehouse-Daten werden immer intensiver zur Steuerung und Unterstützung operativer Customer Management Prozesse eingesetzt, sodass die Qualität der dispositiven Daten eine immer höhere, auch monetäre messbare Bedeutung gewinnt.

Die Datenqualität in Data Warehouse Systemen ist nur durch einen totalen Qualitätsmanagement-Ansatz und durch eine abgestimmte technische Infrastruktur für die Datenqualitätskon-

trolle zu gewährleisten. Datenqualitätsinitiativen starten häufig mit einer rückblickenden Analyse aufgetretener Probleme, identifizieren Brennpunkte und leiten Maßnahmen zur Behebung der speziellen Ursachen ein. Diese Maßnahmen sind wichtig, richtig und kommen doch immer zu spät. Der nächste auftretende Datendefekt kommt mit großer Wahrscheinlichkeit aus einer unerwarteten Quelle, die einzelne Ursache ist im Rückblick betrachtet häufig trivial und niemand möchte sich dafür gerne rechtfertigen. In der Zusammenschau sind diese Ursachen jedoch häufig sehr komplex, sie entstehen zu unterschiedlichen Zeiten und in verschiedenen Prozessen.

House of Quality

Datenqualitätsmanagement muss daher proaktiv und defensiv mögliche Fehlerquellen durch eine hohe Qualität aller beteiligten Softwareentwicklungs- und Betriebsprozesse reduzieren und nach Möglichkeit vermeiden. Erfolgreiches Datenqualitätsmanagement muss sich daher einem „totalen“ Qualitätsmanagement Ansatz zuwenden, der Aufbauorganisation, Prozesse, Regeln und Werkzeuge in einem „House of Quality“ integriert. Die Nachhaltigkeit und Verlässlichkeit der Datenqualität stützt sich auf diese vier Säulen (siehe Bild 1).

Die Ursachen und die Auswirkungen von Datenqualitätsproblemen über-



eines erfolgreichen Datenqualitätsmanagements. Es liegt auf der Hand, dass beispielsweise bereits im Anforderungsmanagement wertvolle Prüffregeln und Testcases gewonnen werden können, die dann im Testprozess aber auch in den regelmäßigen Datenqualitätsprüfungen des operativen Betriebs eingesetzt werden können. Wartungs- und Weiterentwicklungsarbeiten gelingen mit klaren Codierungs- und Designstandards schneller und weniger fehleranfällig. Dies gilt vor allem auch dann, wenn auf Metadaten gestützte Abhängigkeitsanalysen mit einbezogen werden.

Standards und Normen, die gut dokumentiert und allen Beteiligten zugänglich sind, müssen nicht nur für den Softwareentwicklungsprozess definiert werden, sondern auch in Form von Geschäftsregeln, Grenzwerte und Schwankungskorridore für zentrale Kennzahlen, sowie sonstigen Plausibilisierungsregeln für die Entwicklung, den Test und die Qualitätssicherung im laufenden Betrieb etwa mit einem Data Quality-Tool bereitgestellt werden.

Die vierte Säule erfolgreichen Datenqualitätsmanagements sind geeignete, möglichst weitgehend integrierte Werkzeuge und Applikationen vom ETL-

Innovative Ansätze nutzen

schreiten Organisations- und Kompetenzgrenzen in jedem Unternehmen. Daher sind definierte Verantwortlichkeiten für die Datenqualität und vereinbarte, regelmäßige Kommunikationswege zwischen den Betreuern der Quellsysteme und der dispositiven Systeme erfolgskritisch. Die querschnittliche, von mehreren Organisationseinheiten getragene Verantwortung für die Datenqualität bedarf dabei eines klaren Commitments des Top-Managements zur hohen Priorität verlässlicher Daten für das Unternehmen.

Sehr gut definierte und gelebte Prozesse für die Softwareentwicklung und den Betrieb sowie für das Qualitätsmanagement selbst sind das Fundament

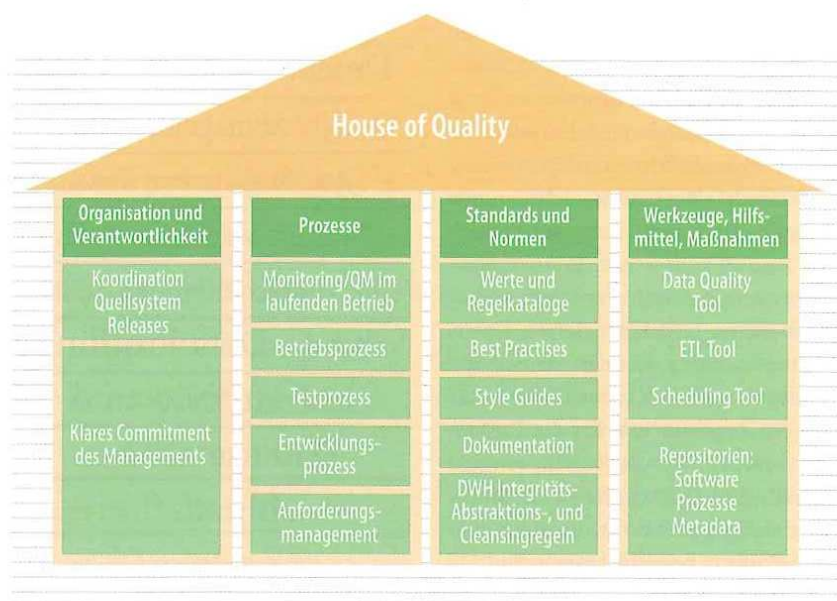


Bild 1: House of Quality.

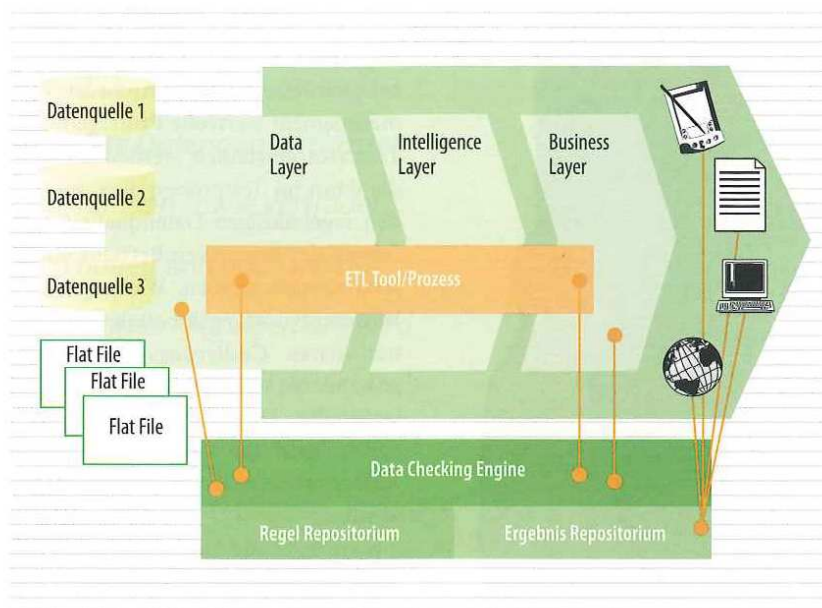


Bild 2: Data Quality-Engine in der DWH-Architektur.

Tool, über die Softwarerepositorien und Metadatenrepositorien bis hin zu Data Quality-Tools. Gute und akzeptierte Werkzeuge erleichtern die Durchsetzung von Softwareentwicklungsprozessen und Standards.

Data Quality Software-Architektur

Es ist offensichtlich, dass Datenqualitätsmanagement daher auch nicht auf den Einsatz einer Spezialsoftware wie ein Data Quality-Tool konzentriert werden kann. Die Vielzahl der Datenbewirtschaftungsprozesse eines Data Warehouse macht eine automatisierte und in die Ladeprozesse integrierte Datenqualitätsprüfung unabdingbar. Die Prüfungen müssen synchron und asynchron zu den ETL-Prozessen entlang der gesamten Datenveredelungsstrecke des DWH ausgeführt werden können, da Datendefekte häufig nicht immer während der Verarbeitung auf Einzelsatzebene erkannt werden können, sondern erst nach dem Ladeprozess in den verarbeiteten und aggregierten Daten ermittelt werden können. Das Data Quality-Framework oder Tool muss über ein Repository für die Prüfregele verfügen, das die Regeln, etwaige Schwell- und Grenzwerte sowie Metadaten speichert und für die automatisierte Prüfung bereithält (siehe Bild 2).

Die Abfrageergebnisse der Prüfungen müssen nach den im Repository

hinterlegten Ergebniserwartungen oder Grenzwerten interpretiert werden. Je nach Schwere eines entdeckten Fehlers wird das Framework eine Benachrichtigung versenden oder einen Workflow



„Innovative Ansätze
wie das Deviation

Detection gehen im Data

Quality Management durch

den Einsatz von Data

Mining-Verfahren noch

einen entscheidenden Schritt

über das Data Profiling

auf Attribut- und

Datensatzebene

hinaus.“

Dr. Jörg Westermayer,
Partner und Practise Leader,
SHS Viveon

anstoßen. Die Prüfergebnisse selbst sollten ebenfalls in einem Repository abgelegt werden, um Nachweise über die Qualität der Datenbewirtschaftungsprozesse führen zu können. Diese Nachweise könnten den Fachanwendern in Form von Datenqualitätsreports regelmäßig als vertrauensbildende Maßnahme mit geringem Aufwand über die bereits genutzte BI-Applikation zur Verfügung gestellt werden.

Multidimensionale Daten stellen eine besondere Herausforderung für ein Data Quality-Framework dar. In OLAP-Reports werden Datenqualitätsprobleme vom Fachanwender in beliebig kleiner Granularität mit einem beliebig geringen Anteil am Gesamtbetrag einer Kenngröße wahrgenommen. Sobald die Provisionszahlung von Data Warehouse-Daten abhängt, sobald für einen Schlüsselkunden der Umsatz für einen bestimmten Abschluss nicht gezeigt wird, leidet die Glaubwürdigkeit des Data Warehouse nahezu unabhängig vom Betrag des Fehlers selbst. Datenqualitätsprobleme werden daher teilweise erst nach der vollständigen Integration und Aggregation in die OLAP-Cubes sichtbar. Prüfregele für besonders sensible Produkt- oder Kundenkennzahlen müssen daher mit multidimensional und hierarchisch ermittelten Ober- und Untergrenzen geprüft werden. Das Framework oder Tool zur Ausführung dieser Regeln sollte gruppierte, multidimensionale und mehrfach parametrisierbare Abfragen unterstützen.

Die Gewinnung der Regeln

Der Erfolg der automatisierten Datenqualitätsprüfung im Data Warehouse steht und fällt mit den eingesetzten Regeln. Die Gewinnung dieser Regeln muss zum einen fester Bestandteil der Anforderungs- und Entwicklungsprozesse sein, zum anderen sollten Data Quality-Tools durch Data Profiling bei der Ermittlung von Prüfregele helfen. Innovative Ansätze des Data Quality-Management ergänzen diese Methoden durch Data Mining-Verfahren.

Die Anwender in den Fachbereichen prüfen die aus dem Data Warehouse generierten Reports und Selektionen vor der Verwendung sehr häufig gegen eine

WEB-TIPP:
www.shs-viveon.com

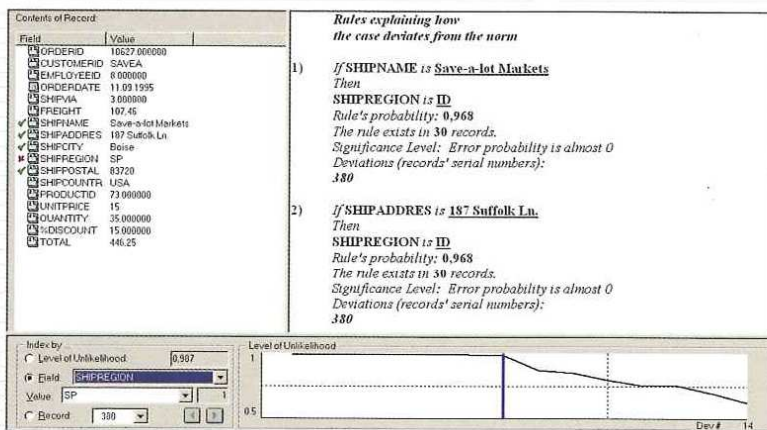


Bild 3: Deviation Detection - Anzeige von Auffälligkeiten.

Vielzahl von Geschäftsregeln und Plausibilitäten. Kennzahlen werden per Augenschein oder in Referenz auf in Excel gerechnete Prüfsummen und Grenzwerte geprüft. Dieses Expertenwissen wird in Data Warehouse Systemen häufig nicht gehoben und proaktiv in die automatisierten Qualitätsprüfungen integriert. Bereits im Anforderungsprozess sollten Geschäftsregeln und Plausibilitätsprüfungen der Fachanwender für die neuen Reports und Kennzahlen abgefragt werden, und für die testgetriebene Entwicklung ebenso eingesetzt werden wie für den Test und die tägliche Qualitätsprüfung nach der Inbetriebnahme. In historisch gewachsenen Reportlandschaften ist es Aufgabe jeder Datenqualitätsinitiative, die Fachanwender systematisch nach Prüfregeln zu befragen und diese in das Regelrepositorium zu übernehmen.

Ähnlich wie die Prüfungen der Fachanwender finden die Tests der Entwickler häufig keinen Eingang in die täglichen, regelbasierten Prüfungen im laufenden Betrieb. Auch hier helfen Entwicklungsstandards und ein einfach mit Plausibilisierungsprüfungen zu befüllendes Regelrepositorium.

Zusätzliche fachliche Prüfungen

Die fachlichen Prüfungen können durch Data Quality-Tools mit Data Profiling-Methoden ergänzt werden. Diese Werkzeuge sind in der Lage, attribut- und datensatzbezogene Muster für Wer-

tebereiche, vorkommende Werte und Verteilungen zu finden und auf Wunsch automatisiert in Prüfregeln zu überführen, die im ETL-Prozess ausgeführt werden können. Ein Verstoß gegen eines der definierten Muster kann dann auf ein Datenqualitätsproblem hinwei-



Multidimensionale Daten stellen eine besondere Herausforderung für ein Data Quality-Framework dar. Prüfregeln für besonders sensible Produkt- oder Kundenkennzahlen müssen daher mit multidimensional und hierarchisch ermittelten Ober- und Untergrenzen geprüft werden.

sen. Data Profiling-Verfahren sind hilfreich im Auffinden von Datendefekten, werden jedoch nicht zuletzt durch die Darstellung der Toolhersteller überschätzt.

Innovative Ansätze im Data Quality Management gehen durch den Einsatz von Data Mining-Verfahren noch einen entscheidenden Schritt über das Data Profiling auf Attribut- und Datensatzebene hinaus. Mit dem Ansatz der „Deviation Detection“ (siehe Bild 3) werden aus ganzen Datenbeständen Irregularitäten und Abweichungen von der Norm toolbasiert durch ungestütztes Lernen extrahiert und in Regeln formuliert. Diese Regeln können im ersten Schritt sehr effizient ohne die Vorarbeit und die unmittelbare Beteiligung der Experten gewonnen werden. Die ermittelten Warnungen und Regeln müssen dann selbstverständlich von den Experten interpretiert und für den Prüfbetrieb freigegeben werden. Die Methode bietet den Vorteil, dass Regeln schneller und vollständiger formuliert werden und Irregularitäten angezeigt werden, sobald sie zum ersten Mal auftreten.

Fazit

Die Sicherstellung der Datenqualität ist einer der entscheidenden Erfolgsfaktoren für die Nutzung von Data Warehouse-Systemen. Mangelnde Datenqualität führt nicht nur zu einer Verringerung der internen Akzeptanz von Data Warehouse-Systemen, sondern kann darüber hinaus wesentlich erhöhte Kosten oder sogar falsche Entscheidungen verursachen.

Die vielen Ausprägungen, Ursachen und Auswirkungen mangelhafter Datenqualität machen deutlich, wie wichtig ein mehrschichtiger auf die Organisation, auf Prozesse, auf Regeln und Werkzeuge abgestimmter Managementansatz für die Gewährleistung einer konstanten Datenqualität für jedes Unternehmen ist. Die Praxis zeigt allerdings, dass nur die wenigsten Unternehmen Prozesse, Methoden und geeignete IT-Infrastrukturen zur Sicherstellung einer angemessenen Datenqualität in ihrer Organisation implementiert haben.

DR. JÖRG WESTERMAYER