

Datenqualität in Kunden- und Materialdatenbanken optimieren

Automatische Datenpflege

Die Qualität von Kunden- und Materialstammdaten lässt in vielen Unternehmen zu wünschen übrig. Datenqualität ist jedoch in fast allen Prozessen und Anwendungen von höchster Bedeutung. **Michael Matzer**

Auf einen Blick

Inhalt
Die Qualität der in Datenbanken gespeicherten Informationen ist in vielen Unternehmen mangelhaft, obwohl korrekte Daten in zahlreichen Prozessen benötigt werden. Der Pforzheimer Datenqualitäts-Spezialist Omikron Data Quality bietet mehrere Werkzeuge an, um die Qualität zu optimieren.

Autor
Michael Matzer, M.A., arbeitet als Journalist, Übersetzer, Rezensent und Buchautor und lebt in der Nähe von Stuttgart.

Zuweilen passiert es beim Briefempfang, dass auf dem Briefkopf zwar der richtige Name steht, aber die Anrede falsch ist – „Sehr geehrter Herr Firma“. Das ist nur ein harmloses Beispiel dafür, wie ein Unternehmen seine Beziehung zu einem potenziellen Kunden durch schlechte Adressqualität belasten kann. In vielen Datenbanken finden sich wesentlich gravierendere Belege für mangelhafte Datenqualität. Laut Umfragen sind rund ein Fünftel aller Adressen in einem Unternehmen fehlerhaft. Ähnlich große Mängel finden sich in Materialstammdaten.

Doch jedes Unternehmen sollte es sich zur Aufgabe machen, seine Datenqualität zu verbessern. Denn die Daten, die in Vertrieb und Marketing, Einkauf und Lager, Buchhaltung und Controlling, Management und IT benutzt werden, bilden ja die Essenz der Prozesse. Stimmt die Qualität dieser Informationen nicht, nützen die besten Prozesse nichts, und die an den Prozessen Beteiligten wundern sich, nicht zuletzt auch die Kunden und Partner.

Die Daten befinden sich in unterschiedlichen ERP-, CRM- und Kundendienst-Anwendungen und sind möglicherweise nicht miteinander verbunden oder abgeglichen. Das führt schnell zu Konflikten: Welche Adresse ist korrekt? Daraus ergibt sich die Notwendigkeit, die Datenqualitätsprüfung in möglichst viele Prozesse zu integrieren. Und je früher man sie in Prozesse integriert, desto weniger Aufwand entsteht und desto geringer ist der Schaden.

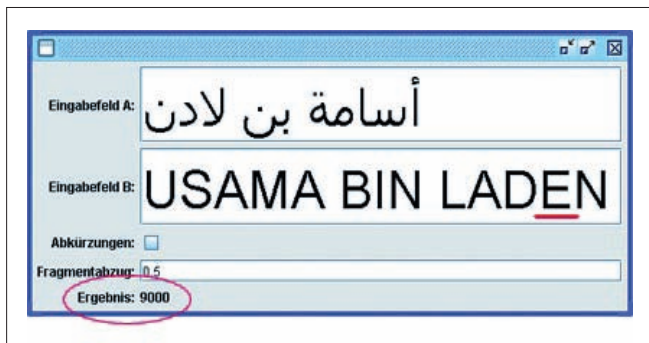
Zu weiteren direkten Vorteilen der Qualitätsoptimierung zählen neben der erwähnten Adressbereinigung auch die Identifikation von Betrügern, die Zuordnung von Produktdaten, die Prüfung von Webanmeldungen und die

Callcenter-Unterstützung mithilfe der fehler-toleranten Adresssuche.

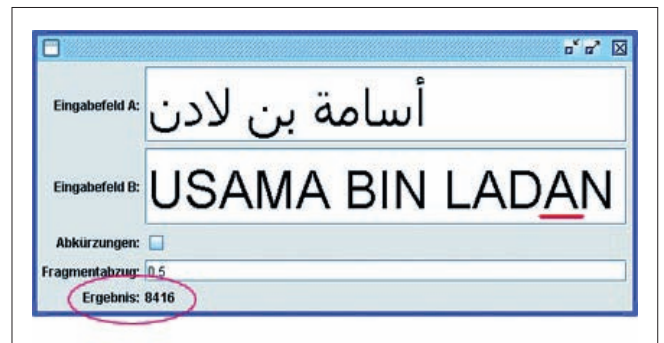
Lösungen für die Verbesserung der Adressqualität sollten die folgenden Grundfunktionen für die Adressbereinigung beherrschen:

- Die Adresse ist an die Post-Schreibweise anzupassen.
- Straßennamen und Hausnummern sind gemäß den Postrichtlinien zu trennen.
- Ein Legal-Form-Extractor sollte die Rechtsform einer Organisation prüfen und gegebenenfalls korrigieren.
- Die korrekte Anredeform für Personen und Firmen ist zu generieren.
- Durchgehend in Großbuchstaben oder Kleinbuchstaben geschriebene Wörter sind in die korrekte Form zu bringen.
- Eine fehlertolerante Adresssuche sollte vorhanden sein.
- Ebenso darf eine erweiterte Adresssuche (Referenzsuche) nicht fehlen.
- Im Datensatz fehlende Informationen, wie etwa die Branche, sollte eine Datenanreicherungs-funktion eruieren.
- Laut EU-Verordnung zur Bekämpfung des Terrorismus muss ein Abgleich mit Sanktionslisten erfolgen.
- Ein Dialog zur Dublettenprüfung sollte helfen, die richtigen Zuordnungen auch bei Abkürzungen und Tippfehlern zu finden.
- Ein Black-Listen-Abgleich sollte die Daten beispielsweise auf bekannte Spaßbesteller, Betrüger und Zahlungsunwillige prüfen. Dazu eignen sich Robinson- oder Nixie-Listen sowie kundeneigene Negativ-Listen.

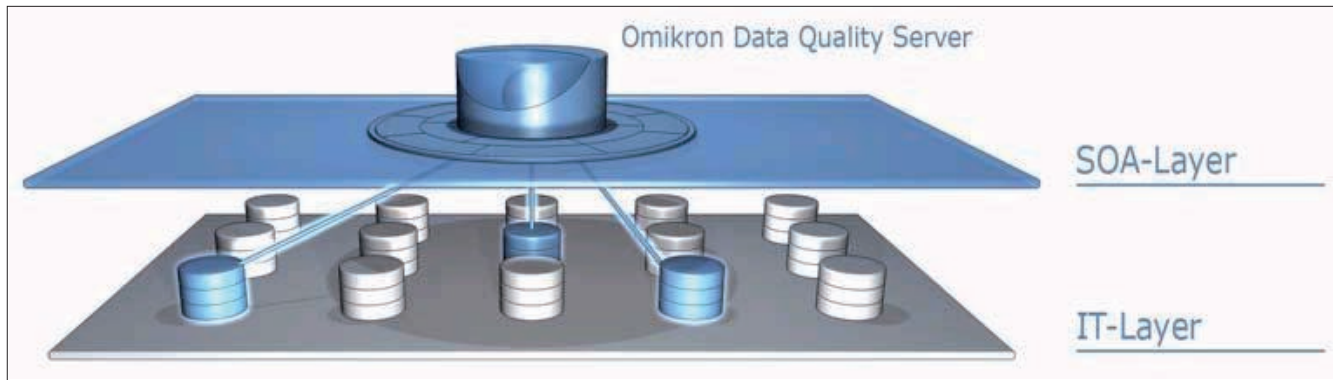
Zu diesen Grundanforderungen kommen fortgeschrittene Funktionen für den Abgleich



Bei der Adressprüfung im Arabischen zeigt Worldmatch die optimale Schreibweise mit dem Indexwert 9000 an (Bild 1)



Suboptimale Schreibweisen erhalten einen niedrigeren Indexwert, hier 8416 (Bild 2)



Der Data Quality Server mit SOA-Layer und den Datenquellen auf dem IT-Layer (Bild 3)

fremdsprachiger Schriften. Deren Bedeutung nimmt zu, je mehr Großunternehmen und Mittelständler ins Ausland expandieren, so etwa nach China und Indien. Doch hier beginnen die ernsthaften Schwierigkeiten, und bei den Produktangeboten zur Datenqualität trennt sich die Spreu vom Weizen.

Ähnlichkeitsprüfung mit FACT und Worldmatch

Kern vieler Datenqualitätsoperationen sind Ähnlichkeitsvergleiche von Daten. Bei der Dublettenprüfung sollen beispielsweise doppelte Adresseinträge gefunden werden – und zwar auch dann, wenn Tippfehler oder Wortumstellungen den Vergleich erschweren. Die eigens von Omikron [1] entwickelten Algorithmen FACT und Worldmatch ermöglichen solche Ähnlichkeitsvergleiche und kommen in vielen Funktionen zum Einsatz.

Das intelligente Ähnlichkeitsverfahren FACT versucht das menschliche Vermögen abzubilden, Ähnlichkeiten erkennen und korrekt zuzuordnen zu können. Kern ist dabei ein spezieller Vergleichsalgorithmus, erweitert um die Omikron-Phonetik. So sind einfache und komplexe Vergleiche im lateinischen Zeichensatz möglich. Der Algorithmus vergleicht zwei beliebige Zeichenfolgen und gibt einen Ähnlichkeitswert zurück. Je ähnlicher die Daten sind, desto höher ist auch der Ähnlichkeitswert, bis zu den maximalen 100 Prozent.

Darüber hinaus bietet FACT Besonderheiten bei Abkürzungen und Hausnummern. So kann beispielsweise die „Adam Müller Hard- und Software GmbH“ der „Hardware + Software-Gesellschaft mbH, A. Müller“ korrekt zugeordnet werden, und die Hausnummer „2-6“ liefert eine Übereinstimmung von 95 Prozent zur Hausnummer „4“.

FACT-Finder, die auf FACT basierende fehler-tolerante Such- und Navigationslösung für Online-Shops, avancierte nach Angaben von Omikron innerhalb von nur zwei Jahren zum deutschen Marktführer im Bereich der intelligenten Produktsuche.

Dublettenfilter mit FACT-Matrix

Die FACT-Matrix dient dem Vergleich eines ganzen Datensatzes in einem Durchlauf. Zu diesem Zweck kann der Nutzer eine Matrix aus verschiedenen Feldern sowie beliebig vielen Dublettentypen zusammensetzen. Er bekommt dann nicht mehr den Ähnlichkeitswert, sondern den Filter genannt, aufgrund dessen die Dublette als ähnlich markiert wurde. Dieses Vorgehen erleichtert die Suche, da der Nutzer die Auswertung der Ähnlichkeitswerte nicht mehr selbst durchführen muss, und erlaubt auch komplexe Dubletten-Definitionen.

Adressoptimierung mit Worldmatch

Bei der Darstellung von nicht-lateinischen Schriftzeichen verlassen sich Datenpfleger heute auf die Zeichenbeschreibungssprache Unicode, die mehr als vier Milliarden verschiedene Zeichen mit 32-Bit-Code abdeckt. Meist reichen jedoch 8- oder 16-Bit-Zeichen. Datenverluste und Fehler treten erst auf, wenn zwischen verschiedenen Alphabeten transkribiert, also in andere Zeichen übertragen wird. Aus dem arabischen Namen „Usama“ wird so in Frankreich „Ousama“, in den USA aber „Osama“ (Bild 1 und Bild 2). Der Grund: Das arabische Alphabet kennt Vokalzeichen nur in Ausnahmen. Die Fehlerquellen in internationalen Adressbeständen können zusätzlich durch Buchstabendreher und Tippfehler anwachsen. Bei der Zusammenführung von Adressdatenbanken, beispielsweise nach Firmenfusionen, treten zudem in vielen Fällen Inkonsistenzen auf, weil zwei verschiedene Erfassungsregeln miteinander in Konflikt geraten.

Die Technologie Worldmatch von Omikron versetzt Unternehmen in die Lage, Schriftzeichen unterschiedlicher Sprachen direkt miteinander zu vergleichen, und zwar ohne sie dabei in Unicode umzurechnen. So können bei einem Dubletten-Abgleich beispielsweise lateinische Adressen mit Adressen aus den arabischen oder asiatischen Alphabeten miteinander verglichen werden. ▶

„Worldmatch ist das erste multidimensionale Abgleichprogramm, das in der Lage ist, alle Datenähnlichkeiten weltweit auf einmal zu erfassen“, erklärt Omikron-Geschäftsführer Carsten Kraus. „Möglich wird das erst durch den intelligenten Assoziativ-Algorithmus, der ganz unterschiedliche Zeichensätze miteinander in Bezug setzen kann. Dabei werden sogar die Eigenheiten der verschiedenen Schriften und Sprachen berücksichtigt, so etwa auch die optischen Ähnlichkeiten zwischen chinesischen Zeichen.“

Der Data Quality Server

Die zugrunde liegenden Algorithmen sind beim Hersteller Omikron in seinen Produkten FACT, FACT-Finder, AdressCenter und Worldmatch zu finden. Sie sind mit der serviceorientierten

Serverplattform Data Quality Server (DQS) gekoppelt (Bild 3), die ihre Funktionen als Web-Services in die Geschäftsprozesse einbindet. Datenqualität sollte in die Geschäftsprozesse eines Unternehmens eingebunden werden, um den größten Kostenvorteil zu bieten. „Je später die Fehlerbereinigung erfolgt, desto aufwendiger und teurer wird sie auch“, gibt Carsten Kraus zu bedenken.

Der DQS ist als programmierbare Serverplattform für eine serviceorientierte Architektur (SOA) ausgelegt, lässt sich also mit SOA-Technik integrieren und mit unterschiedlichen Programmiersprachen ansprechen. Das mitgelieferte Management Studio bietet eine grafische Benutzeroberfläche, mit welcher der Administrator alle Optionen zu Datenverbindungen sowie Funktionen und Benutzer-Rechte festlegen

Ein Praxisbeispiel

Im Folgenden wird der Server für die Adressprüfung anhand des Moduls FACT-Finder konfiguriert, um eine Adresssuche zu ermöglichen. Im Vor-Projekt erfolgen die Planung der Zielstruktur sowie die erste Datenladung aus dem bereits vorhandenen Datenbestand. Dieser Schritt ist meist ein iterativer Vorgang, bei dem die Daten analysiert und im Dialog bereinigt werden. Zusätzliche Erkenntnisse aus der Analyse und Bereinigung der Daten können hier noch in die zu definierenden Prozesse einfließen (Bild 6).

Im prozessorientierten Teil definiert der Nutzer, wie Datenqualitätsprüfungen für die einzelnen Geschäftsprozesse realisiert werden können und sollen. Dabei wird unterschieden zwischen

- allgemeinen DQ-Prozessen, etwa welche Prüfungen neue Kunden in der zentralen Adressdatenbank durchlaufen müssen, und
- speziellen DQ-Prozessen, etwa bei einer Bestellung durch Partner.

Der Workflow für das Praxisbeispiel beschreibt den Einsatz des DQ-Servers im Rahmen einer Data Quality Firewall, die verschiedene automatisierte Prüfungen durchführt: Es wird die postalische Richtigkeit der Adressen sichergestellt und

die Dublettenprüfung gegen die bestehende Datenbasis durchgeführt.

Zunächst ist die Data Quality Firewall einzurichten. In der Activity *DQServer_Login* werden die Zugangsdaten zum DQ-Server hinterlegt. Dazu sind der URL, unter der der Server erreichbar ist, sowie Benutzername und Passwort nötig, mit denen der Serverzugriff erfolgen soll (Bild 7).

Die postalische Prüfung in der Activity *Postal_Correction* benötigt zum Aufruf lediglich die Adressbestandteile, die für die Adressprüfung benötigt werden, sowie den Namen der auf dem Server hinterlegten Einstellungen. In den Servereinstellungen der postalischen Prüfung kann der Nutzer im Management Studio beispielsweise die Korrekturgenaugigkeit sowie weitere

Parameter wie den Umgang mit Umlauten oder der Groß-Kleinschreibung, aber auch die zu verwendende Referenzdatenbank festlegen. So kann er mehrere Einstellungen und Referenzdatenbanken auf dem Server parallel verwenden. Diese Funktion liefert die bereinigten Adressbestandteile sowie Statusinformationen, die als Entscheidungsgrundlagen dienen. Danach kann eine Prüfung der

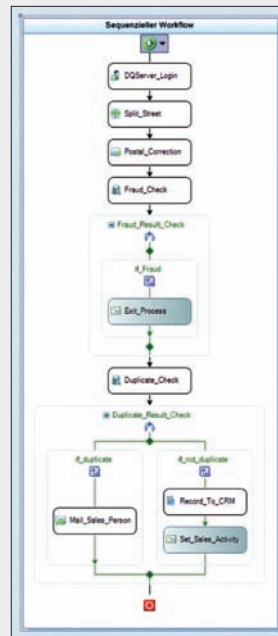
Daten gegen die Betrugsdatenbasis erfolgen.

In der Duplikatsprüfung *Duplicate_Check* wird der bereits als Kunde klassifizierte Adressdatensatz zu den bestehenden Kunden hinzugefügt. Ziel ist es, eine in der Regel ungewollte Doppelanlage der Unternehmenskunden von Beginn an zu verhindern. Dieser Schritt ist immens wichtig, denn die Doppelanlage lässt sich nur schwer rückgängig machen.

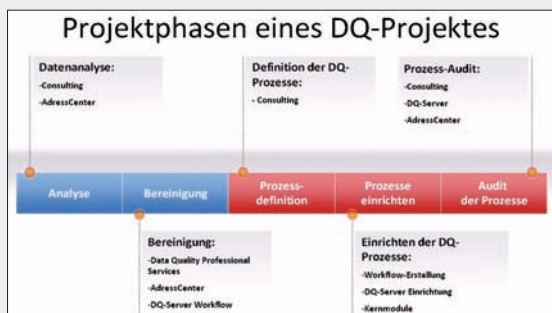
Zunächst benötigt man den Zugriff auf die Kundendaten – siehe oben. Die noch fehlenden Abgleicheinstellungen kann der Nutzer mittels vorkonfigurierter Templates einrichten.

Das Ergebnis der Duplikatsprüfung wird in der Activity

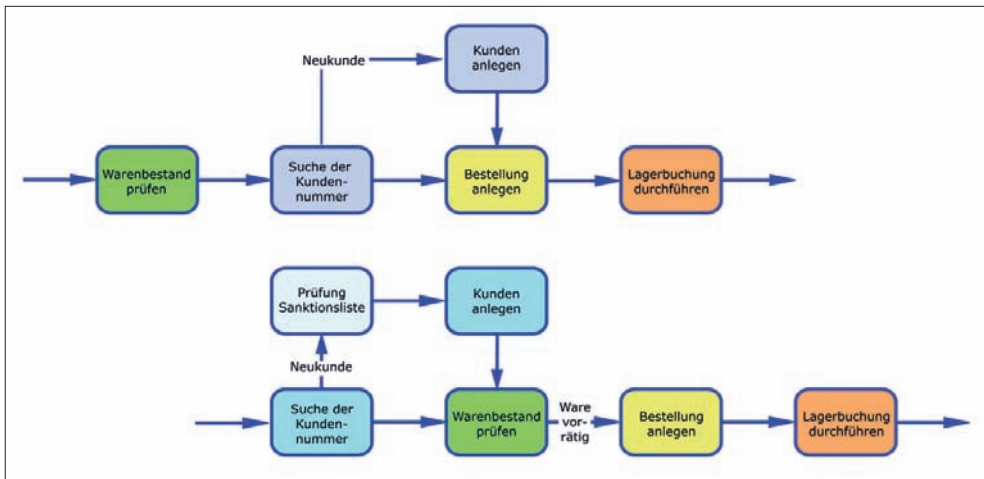
Duplicate_Result_Check ausgewertet. Im Beispiel wird nur zwischen einem Treffer und einem Nicht-Treffer unterschieden, der mögliche Bereich „Unsichere Treffer“ wird den Treffern zugeschlagen. In letzterem Fall wird eine Mail an einen entsprechenden Empfänger versandt (Activity *Mail_Sales_Person*), mit dem Hinweis, welchem bereits existierenden Ansprechpartner die Daten zugeordnet wurden. Handelt es sich um eine Neuanlage, wird in *Record_To_CRM* der Datensatz an das CRM-System gesandt und der Nachfolgeprozess eingerichtet (*Set_Sales_Activity*). Damit steht der Workflow auch schon, und die Data Quality Firewall ist einsatzbereit.



Mit der Workflow-Engine des DQS können einfache und komplexere Prozesse erstellt werden (Bild 7)



Die Phasen eines Projektes mit dem DQ-Server von Omikron. Blau ist das Vor-Projekt, rot der prozessorientierte Teil (Bild 6)



Der DQS mit SOA als übergreifender Architektur ermöglicht Agilität in den täglichen Arbeitsprozessen (Bild 4)

kann. Danach kann der Benutzer eigene Datenqualitätsprozesse definieren (Bild 4) und die entsprechenden DQS-Funktionen innerhalb der Prozesse aufrufen.

Der Data Quality Server wird auf einem leistungsstarken Windows-System ab XP beziehungsweise Windows Server 2003 installiert. Ebenso erforderlich sind die Microsoft Internet Information Services (IIS) ab Version 5.0 und das Microsoft .NET Framework 2.0. Letzteres wird mit dem DQ-Server implementiert und umfasst auch ASP.NET 2.0.

Damit der DQS eingerichtet werden kann, müssen zuerst die Internet Information Services konfiguriert werden. Die Hilfefunktion unterstützt den Nutzer bei der Einrichtung und beim Verbinden des Servers mit dem Management Studio. Hier findet der Anwender alle Optionen zur Einrichtung von Datenverbindungen.

Der DQS sorgt für die Integration einer Vielzahl von wählbaren Datenquellen. Externe Datenquellen lassen sich über Standard-Datenbankschnittstellen ansprechen und nutzen, ganz gleich, ob Microsoft SQL Server, Sun MySQL, Oracle 11g, Microsoft Access oder individuelle Datenbankformate. Der Server unterstützt generell OLE DB sowie spezielle Treiber (ADO.NET).

keit verringert sich deutlich mehr, als wenn auf einer Datenbank in einem der internen DQS-Formate gearbeitet würde. Das wirkt sich besonders dann aus, wenn sehr viele Anfragen (etwa Suchanfragen) gleichzeitig abgeschickt werden. Für eine optimale Performance müssen Indizes auf der Datenbank vorhanden sein oder angelegt werden, die denselben Aufbau wie die dazugehörige Such-Umgebung besitzen.

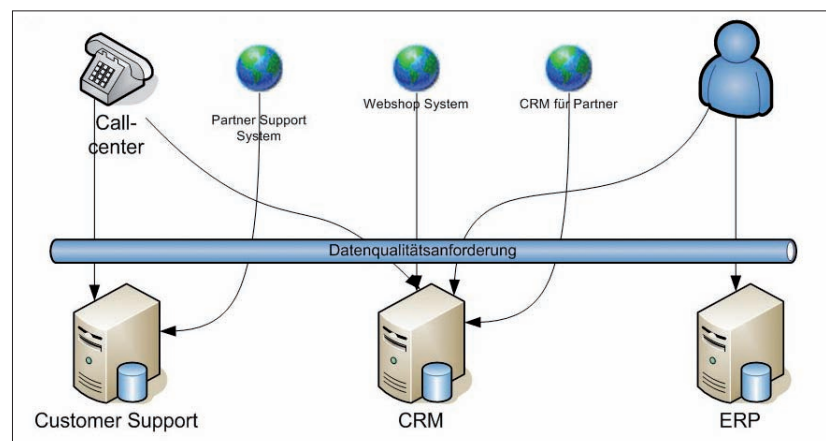
Bei der Verwendung eines synchronisierten Datenbestandes wird hingegen eine Kopie der Live-Datenbank angefertigt, auf der Datensätze bearbeitet oder gesucht werden. Kopiert man beispielsweise die Datenbank eines CRM-Systems, werden neue Einträge auf der Live-Datenbank abgespeichert; gesucht wird aber auf der Datenbankkopie des DQ-Servers, was dann schneller erfolgen kann. Damit die Such-Tabelle aktuell bleibt, muss sie in regelmäßigen Abständen mit den Originaldaten synchronisiert werden. Auf Wunsch kann dies automatisch durch den DQ-Server erfolgen. Neue Datensätze lassen sich mit kurzer Wartezeit finden. Das Arbeiten mit einem synchronisierten Datenbestand ist damit wesentlich performanter als beim oben genannten Direktzugriff (Bild 5). [bl]

[1] Omikron Data Quality GmbH; www.omikron.net

Zwei Wege der Datenintegration

Es gibt zwei Möglichkeiten, die DQS-Module mit den eigenen Daten in Verbindung zu bringen. Die erste Variante besteht darin, direkt auf den Originaldaten zu arbeiten. Die zweite Möglichkeit sieht vor, dass der Anwender den Datenbestand dupliziert und auf den Tabellen arbeitet, die der DQS verwaltet. Beim Direktzugriff werden die DQS-Funktionen auf die bereits bestehende Datenbank angewendet. Der Vorteil besteht darin, dass dann die Datenbank immer aktuell ist.

Die Nachteile sind jedoch: Die Datenbank wird zusätzlich belastet, und die Geschwindigkeit



Der DQS führt die notwendige Vereinheitlichung der Datenqualität in den Systemen des Unternehmens herbei (Bild 5)