

Lösungen für Customer Intelligence,
Customer Communications und Customer Care.

Alternative Ansätze der Datenintegration – für das Wert- und Qualitätsmanagement

Ein strukturiertes Vorgehen zur Einbindung von Datenqualitätsservices
in den Datenintegrationsablauf

WHITEPAPER:

DATENQUALITÄT UND DATENINTEGRATION

David Loshin • President
Knowledge-Integrity, Inc.



Alternative Ansätze der Datenintegration – für das Wert- und Qualitätsmanagement

Ein strukturiertes Vorgehen bei der Einbindung von Datenqualitätsservices bei den Datenintegrationsablauf

2

ZUSAMMENFASSUNG

ANGESICHTS DER DRAMATISCH STEIGENDEN DATENFLUT VERLASSEN SICH IMMER MEHR ERFOLGREICHE UNTERNEHMEN AUF IHRE REPORTING- UND ANALYSELÖSUNGEN, UM DIE ART UND WEISE, WIE SIE GESCHÄFTE TÄTIGEN, AKTIV ZU VERBESSERN. WÄHREND JAHRZEHNTELANGE INVESTITIONEN IN TRANSAKTIONS- UND ABLAUFORIENTIERTE GESCHÄFTSANWENDUNGEN ZUR ENTSTEHUNG ISOLIERTER DATENINSELN GEFÜHRT HABEN, GIBT ES NUN EINE WACHSENDE ZAHL VON UNTERNEHMENSAPPLIKATIONEN, BEI DENEN AUF DATENSÄTZE AUS VERSCHIEDENEN QUELLEN ZUGEGRIFFEN WERDEN MUSS. BEISPIELE HIERFÜR SIND ANWENDUNGEN AUS DEN BEREICHEN BUSINESS INTELLIGENCE UND DATA WAREHOUSING, CUSTOMER RELATIONSHIP MANAGEMENT UND ENTERPRISE RESOURCE PLANNING SOWIE BESONDERS KOMPLEXE BUSINESS ANALYTICS-APPLIKATIONEN.

Neuer Kommunikationskanal zum Kunden Datenintegration ist allgegenwärtig

Angesichts der dramatisch steigenden Datenflut entscheiden sich immer mehr Unternehmen für Reporting- und Analyselösungen, um die Art und Weise, wie sie Geschäfte tätigen, aktiv zu verbessern. Dabei wird die Datenzentralisierung zum Schlüssel für die Einführung strategischer Unternehmensanwendungen. Operational Data Stores, Data Warehouses, Data Marts, Mash-ups, angebundene Fachanwendungen, Self-Service Reporting, Datenaustausch und andere analytische und operative Anwendungen erfordern ein sehr viel höheres Maß an gemeinsamer Datennutzung, als dies in der Vergangenheit der Fall war.

Den Informationsbedarf dieser nachgeordneten Geschäftsanwendungen zu befriedigen, wird somit zu einer Schlüsselaufgabe. In der Praxis bedeutet dies, dass Daten von ihren Ausgangsquellen zu den Zieldatensystemen »verschoben« werden müssen. Anders ausgedrückt: Zur Stillung des Informationsbedarfs ist die Datenintegration zwingend erforderlich.

Alternative Ansätze bei der Datenintegration

Mit steigender Komplexität der von Endnutzern gestellten Anforderungen entstanden unterschiedliche Ansätze für die Datenintegration. Beim herkömmlichen ETL-Ansatz (Extract, Transform, Load) werden die Daten aus der jeweiligen Quelle extrahiert und in einem »Staging Area« genannten Datenbereich zwischengespeichert, in dem sie verarbeitet und in eine Zieldarstellung umgewandelt (transformiert) werden. Ein alternativer Ansatz besteht in der Datenvirtualisierung, bei der die Daten in der Quelle gespeichert bleiben und bei Bedarf eine Konzeptansicht erzeugt wird.

Herkömmlicher ETL-Prozess

Der gängigste Ansatz für die Datenintegration besteht in einer mehr oder minder abgewandelten Form des ETL-Prozesses. Da die Datenquellen für End- und nachgeordnete Geschäftsanwendungen häufig ganz unterschiedliche Formate, Datei- und/oder Tabellenstrukturen aufweisen und mitunter sogar verschiedene Zeichenkodierungen verwenden, werden Datensatzrepräsentationen in der Regel normalisiert, bevor sie in einem End-Zielsystem zusammengeführt werden. Dabei benötigt die Datenextraktionskomponente spezielle Routinen für den Abruf der Daten aus ihren jeweiligen Quellen. Dies wiederum erfordert spezielle Transformationsschritte, die eine Reihe von Funktionen für die Normalisierung, Bereinigung, Standardisierung, Ableitung, Übersetzung usw. der Daten anwenden, um diese in ein Format zu bringen, das mit anderen transformierten Datenquellen so weit übereinstimmt, dass eine Aufnahme in die Zieldatensysteme möglich ist.

Erst zu diesem Zeitpunkt können die Daten verteilt und in das Zielsystem geladen werden, wobei sie die vorhandenen Daten entweder überschreiben oder den bestehenden Datensatz in regelmäßigen Abständen ergänzen.

Datenvirtualisierung

Im Gegensatz zum herkömmlichen Ansatz, bei dem Daten aus mehreren Quellen extrahiert und in einer Staging Area zwischengespeichert werden, können die Quelldaten bei einem als Datenförderer oder Datenvirtualisierung bezeichneten Verfahren an Ort und Stelle verbleiben.

Die Datenvirtualisierung schafft Abstraktionsebenen für eine Vielzahl von nativen Datenquellen und stellt nebenbei relationale Ansichten bereit, ohne dass die Daten hierfür aus der jeweiligen Quelle extrahiert werden müssen. Dieser Ansatz ermöglicht die Erzeugung wiederverwendbarer Datenservices, während die üblicherweise in einer Virtualisierungsumgebung implementierten Abstraktionsebenen eine standardisierte, logische Darstellung von Unternehmensdatenkonzepten ermöglichen. Damit steht einer Vielzahl von

Alternative Ansätze der Datenintegration – für das Wert- und Qualitätsmanagement

Ein strukturiertes Vorgehen bei der Einbindung von Datenqualitätsservices bei den Datenintegrationsablauf

4

unterschiedlichen Downstream-Nutzern eine Datenansicht zur Verfügung, die sowohl in struktureller als auch in semantischer Hinsicht konsistent ist.

Herausforderungen an die Datenqualität: Vollständigkeit, Konsistenz, Angemessenheit

Mit steigendem Interesse an der Entwicklung von Geschäftsanwendungen, die primäre Datenquellen für sekundäre Einsatzzwecke wiederverwendbar machen, entwickeln sich oft auch weit auseinanderklaffende Erwartungen der Endnutzer an die Daten. Dies gilt besonders hinsichtlich der Datenqualität. Wenn jedoch Datensätze für Zwecke verwendet werden, für die sie ursprünglich nicht vorgesehen waren, muss ihre Bedeutung von den Nutzern häufig neu definiert und interpretiert werden.

Das wiederholte und immer wieder unterschiedliche Extrahieren und Transformieren der Daten für verschiedene Anwendungen kann zu uneinheitlichen Ergebnissen und ständig notwendigem Datenabgleich führen. Das Resultat sind ein geringeres Vertrauen in die Daten, unnötiger Zeit- und Arbeitsaufwand sowie fragwürdige Ergebnisse. Einige der häufigsten Probleme lauten:

- Fehlende Datenelementwerte, wodurch Zähl- und andere aggregierte Werte verzerrt werden
- Erhebliche Schwankungsbreiten bei anerkannten Bezugsdaten, die zu Inkonsistenzen und Ungenauigkeiten führen
- Unterschiedlicher Formate, Strukturen und semantische Sichten auf die Daten durch End-Geschäftsanwendungen, in deren Folge die Gefahr besteht, dass aus vergleichbaren Analysen verschiedene Schlussfolgerungen gezogen werden
- Inkonsistenzen beim Reporting, wodurch die erzeugten Berichte laufend abgeglichen werden müssen

- Anwendung verschiedenartiger semantischer Systeme, woraus Fehlinterpretationen und unerfüllte Erwartungen an die Angemessenheit der Daten entstehen

Der Mangel an Standards für Strukturen, Formate und Definitionen wiederverwendeter Daten äußert sich im Wesentlichen in Problemen, die sich aus unregelmäßigen Datenintegrationsprozessen ergeben. Typische Probleme sind unvollständige Daten, inkonsistente Daten, sowie Daten, welche die Erwartungen an ihre Angemessenheit nicht erfüllen.

Steigerung des Datenwerts durch geregelte Datenqualitätsservices

Probleme, die infolge der Datenintegration auftreten, können gelöst werden, indem sie zunächst als solche wahrgenommen und definiert werden. Anschließend muss der Integrationsprozess so umgestaltet werden, dass Datenfehler behoben werden, bevor sie größeren Schaden anrichten können. Glücklicherweise kann eine geregelte Herangehensweise an die Einbindung von Datenqualitätsservices in die Datenintegration eine Reihe der Probleme lösen, die infolge einer unkontrollierten Datenkonsolidierung auftreten können. Dabei besteht der erste Schritt darin, eine Reihe von zentralen Data Governance-Prozessen einzuführen, mit denen die Erfassung und Standardisierung der an die Qualität von Unternehmensdaten gestellten Anforderungen definiert wird. In einem zweiten Schritt werden die Verfahren für das Datenqualitätsmanagement in die Datenintegrationsstrategie eingebettet.

Praktiken für Data Governance

Während ein Data Governance-Programm eine Vielzahl von Datenmanagementprozessen umfasst, lassen es die praktischen Erfordernisse der Datenintegration angemessen erscheinen, sich auf eine Auswahl dieser Praktiken zu beschränken. Dabei handelt es sich um Praktiken, welche die Datenintegration unmittelbar unterstützen, wie beispielsweise:

- **Analyse der Datenanforderungen** – Bei der Entwicklung von Geschäftsanwendungen wird der detaillierten Beschreibung der Datenanforderungen meist für weniger wichtig erachtet als die Analyse der Funktionsanforderungen. Da Enterprise-Projekte wie Data Warehousing oder Customer Relationship Management jedoch die Grenzen einzelner Geschäftsbereiche überschreiten, wird ein sorgfältig definierter Prozess benötigt, um die Datenanforderungen aller End-Nutzer zu erfassen, zu dokumentieren und zusammen zu bringen und diese Erwartungen dann in Datenanforderungen umzumünzen, die auf alle relevanten Datenquellen angewendet werden. Dies bedeutet nicht nur ein radikales Umdenken bei der Anforderungserfassung, sondern erfordert auch eine breit angelegte Übersicht, wie sie von einer Data Governance-Infrastruktur bereitgestellt wird.
- **Überprüfung der Datenstandards** – Die Festlegung von Datenstandards kann das Problem der mangelnden Konsistenz lösen, indem insbesondere die Definitionen und die Semantik der Datenelemente aufeinander abgestimmt werden. Durch die Beteiligung wichtiger Stakeholder aus dem gesamten Unternehmen an der Überprüfung und Genehmigung der geplanten Datenstandards entsteht Vertrauen und die Überzeugung, dass die Standards sämtliche Anforderungen der End-Nutzer auch wirklich berücksichtigen.
- **Metadatenmanagement** – Hierzu gehören Prozesse, die die genehmigten Datenstrukturen und -definitionen für Referenz-Datendomänen und den Datenaustausch dokumentieren und ein Instrument für die Kommunikation dieser Standards bereitstellen.

Datenqualitätsservices

Data Governance-Praktiken vereinfachen die Einführung von Verfahren zur Sicherstellung der Datenqualität, wie z. B.:

- **Analyse und Standardisierung** – Analyse bezeichnet hier einen Prozess, der ausgehend von vorgegebenen Formaten, Mustern und Strukturen ermittelt, ob bestimmte Datenwerte definierten Vorgaben entsprechen. Dieser Prozess kommt gemeinsam mit Standardisierungsregeln zum Einsatz, mit denen die Eingangsdaten in eine Form gebracht werden, die effektiver verwendet werden kann, entweder um die Darstellung zu standardisieren (sofern es sich um eine gültige Darstellung handelt) oder um die betreffenden Werte zu korrigieren (sofern bekannte Fehler festgestellt werden). Analyse und Standardisierung können eine Bibliothek von Datendomänen und Regeln nutzen, um die Datenwerte in mehrere Einzelkomponenten aufzuteilen und diese Komponenten dann in ein normiertes Format zu bringen. Mithilfe der Standardisierung können außerdem Abkürzungen in ganze Wörter und Nicknames in Standardnamen umgewandelt werden; auch die Übersetzung von einer Sprache in die andere (z. B. vom Deutschen ins Englische), die Korrektur von Rechtschreibfehlern und die Verringerung der Werteverianz ist möglich, um eine bessere Datensatzverknüpfung für die De-Duplizierung und Datenbereinigung zu erzielen.
- **Datenbereinigung** – Wenn Datenwerte inkonsistent oder fehlerhaft sind und die Fehler nicht am Datenursprung behoben werden können, besteht auch die Möglichkeit, Transformationsregeln anzuwenden. Mit diesen Regeln lassen sich Datenwerte zuschreiben, Namen oder Adressen korrigieren, irrelevante und/oder bedeutungslose Daten löschen und sogar doppelte Datensätze zusammenführen. Mit der direkten Bereinigung der Daten wird sichergestellt, dass diese ein bestimmtes Tauglichkeitsniveau einhalten. Wenn dieser Ansatz im Rahmen der Datenintegration auf Analyse, Standardisierung und Bereinigung angewendet wird, werden die Transformationsvorgänge so standardisiert, dass allen End-Nutzern eine in sich schlüssige Ansicht der Daten bereitgestellt wird.

Alternative Ansätze der Datenintegration – für das Wert- und Qualitätsmanagement

Ein strukturiertes Vorgehen bei der Einbindung von Datenqualitätsservices bei den Datenintegrationsablauf

6

- Datenvalidierung – Wenn keine Koordination zwischen den verschiedenen Datennutzern gegeben ist, wenden diese unter Umständen ein und dieselben Datenvalidierungen auf unterschiedliche Arten an. Selbst wenn sie dieselben oder ähnliche Regeln anwenden, ist es in der Praxis äußerst unwahrscheinlich, dass diese Regeln in derselben Reihenfolge ausgeführt werden oder dass die zulässigen Grenzwerte identisch sind. Daraus ergibt sich, dass selbst bei ein und derselben Quelle das Ergebnis der Validierung ganz unterschiedlich ausfallen kann.

Indem der Datenintegrationsprozess um einen standardmäßigen Satz von Datenvalidierungen ergänzt wird, können Beschränkungen an konkreten Stellen des Informationsflusses getestet werden. Dies senkt die Gefahr von Inkonsistenzen.

Die Erfassung der von den End-Nutzern gestellten Anforderungen an die Datenqualität ermöglichen die Formulierungen von Datenqualitätsregeln. Wenn die Einhaltung dieser Regeln an einer frühen Stelle des Prozesses überprüft wird, kann dies dazu beitragen, dass die Datenqualität den Geschäftsanforderungen entspricht und dass mögliche Probleme frühzeitig erkannt und konsistent gelöst werden.

Wichtige Überlegungen

Der Ansatz für die Umsetzung der Datenintegration sollte nicht im Widerspruch zu dem Wunsch stehen, die Qualität der betreffenden Daten auf eine kohärente und in sich stimmige Art und Weise zu verbessern. Die meisten Anbieter traditioneller ETL-Werkzeuge haben die Notwendigkeit für Datenqualitätsverfahren erkannt, und viele sind Partnerschaften mit Anbietern von Tools zur Sicherung der Datenqualität eingegangen bzw. haben entsprechende Anbieter in ihr Unternehmen eingegliedert.

Mittlerweile findet sich kaum noch ein Komplettlösung für das Extrahieren, Transformieren und Laden von Daten, das auf die Definition und Anwendung integrierter Analyse-, Standardisierungs-, Bereinigungs- und Validierungsschritte verzichtet.

Alternativ hierzu werden Datenvirtualisierungstools auch vermehrt mit Tools und Technologien zur Sicherstellung der Datenqualität verknüpft. Die mit der Datenvirtualisierung einhergehende Abstraktion bietet Datenmanagementteams eine wertvolle Möglichkeit, ihre End-Datennutzer aktiv einzubinden, Anforderungen an die Datenqualität zu erfragen und Datenattribut-basierte Validierungen direkt in eine der Abstraktionsebenen einzubetten. Die Konsolidierung der Anforderungen an die Datenqualität und die Implementierung von Validierungsbeschränkungen an konkreten Stellen des Datenproduktionsflusses verringert die Gefahr von Inkonsistenzen, trägt dazu bei, dass die Datenqualität den Anforderungen der End-Nutzer entspricht und weist die Data Stewards auf potenzielle Probleme hin, die somit frühzeitig im Datenintegrationsprozess erkannt und behoben werden können.

Data Governance-Praktiken vereinfachen die Einführung von Verfahren zur Sicherstellung der Datenqualität.

Hier die wichtigsten Aussagen in Kürze: Datenintegration ist heutzutage ein Thema, das im gesamten Unternehmen von größter Relevanz ist. Durch die Einführung von geregelten Prozessen, welche die Wiederverwendung von Daten auf konsistente Art und Weise vereinfachen, steigt das Vertrauen in Reporting und Analyse, was Stakeholdern unternehmensweit zu Gute kommt.

Ausführliche Informationen zu Lösungen für Datenqualität und Datenintegration erhalten Sie telefonisch oder auf den Websites von Pitney Bowes Business Insight.

Über den Autor

David Loshin ist President von Knowledge Integrity, Inc., einer Beratungs- und Entwicklungsfirma, die maßgeschneiderte Lösungen für das Informationsmanagement sowie Beratungsservices für derartige Lösungen, Schulungen zum Thema Informationsqualität und Geschäftsregellösungen anbietet. Loshin ist Autor von »Master Data Management, Enterprise Knowledge Management – The Data Quality Approach and Business Intelligence – The Savvy Manager's Guide« und hält häufig Vorträge zur Maximierung des Informationswerts. Sie erreichen ihn unter loshin@knowledge-integrity.com oder telefonisch unter +1 301 754-6350.

USA

Pitney Bowes Business Insight
4200 Parliament Place, Ste 600
Lanham, MD 20706-1844
Tel: +1 301-731-2300
Fax: +1 301-731-0360
www.pbinsight.com

EUROPA

Pitney Bowes Business Insight
Minton Place
Victoria Street
Windsor, Berkshire SL4 1 EG
Tel: +44 (0)1753 848200
Fax: +44 (0)1753 621140
www.pbbusinessinsight.co.uk

ASIEN/AUSTRALIEN

Pitney Bowes Business Insight
Level 7, Elizabeth Plaza
North Sydney, NSW 2060
Tel: 61 2 9437 6255
Fax: 61 2 9439 1773
www.pbinsight.com

DEUTSCHLAND/ÖSTERREICH/SCHWEIZ

Pitney Bowes Software Europe GmbH
Pitney Bowes Business Insight Division
Grafinger Straße 2
81671 München
Tel.: +49 (0)89 462387-0
Fax: +49 (0)89 462387-44
www.pbinsight.de

PITNEY BOWES BUSINESS INSIGHT
IS HEADQUARTERED
OUT OF THE UNITED STATES.