

# Wie Sie aus Ihrem Data Warehouse *die Wahrheit* holen

*Data Warehousing hat sich zur grundlegenden Basis für Business-Intelligence-Prozesse entwickelt. Es beinhaltet eine dokumentierte Struktur und ermöglicht die Konsistenz analytischer Anwendungen. Je breiter analytische Methoden und Werkzeuge eingesetzt werden, desto entscheidender wird der Data-Warehousing-Prozess. Neue Metadaten-Services liefern Endanwendern eine „einzige Version der Wahrheit“.*

Data Warehousing ist mehr als ein ETL- (Extrahieren, Transformieren und Laden) Prozess für eine Datensammlung. Wichtig ist v. a. die Integration der Infrastruktur dieser Ebene. Dieser Artikel untersucht die Bedeutung der Metadatenverarbeitung und wie eine Modernisierung von Data Warehouses durch zwei neue Services – Multilevel Planning/Administration und Data Quality Integration – möglich wird, welche Effizienz und Verwaltbarkeit von ETL-Prozessen stark beeinflussen.

## **Die Bedeutung des Metadaten Processing**

Wie viele andere IT-Abläufe auch, basiert Data Warehousing auf Metadaten – Informationen, die die zu Grunde liegenden Daten beschreiben –, sowie auf Prozessen, die basierend auf diese Metadaten ausgeführt werden. Bei der Erstellung von Infrastrukturplänen für grosse Data Warehouses, müssen die Anforderungen an Metadaten als kritischer erster Schritt im Planungsprozess betrachtet werden.

ETL-Server der nächsten Generation werden in hohem Masse von Metadaten-Services abhängen. Solche Services können als Respository betrachtet werden, welches in andere Repositories mit bestimmten Funktionen eingegliedert wird. So erfordert z. B.

eine typische Data-Warehousing-Methodologie ein Modell mit drei getrennten Ebenen Entwicklung (DEV), Test (TEST) und Produktion (PROD). Jede Ebene wird durch eine Anzahl von Repositories repräsentiert, welche den Data Warehouse-Betrieb für diese Ebene umfassen. Innerhalb einer Ebene, wie z. B. der DEV-Umgebung, können mehrere ETL-Entwickler an verschiedenen Aspekten eines grösseren Projekts arbeiten, wobei jedes dieser Projekte von der Unabhängigkeit der Metadaten während der anfänglichen Entwicklungsphase profitiert. Ein wichtiger Aspekt innerhalb der Metadatenverarbeitung ist das Change Management, das eine Trennung unabhängiger Projekte erlaubt, um paralleles Arbeiten zu ermöglichen. Ohne Change Management kann eine unbedachte Änderung ein Projekt zum Stillstand bringen, da jeder Anwender von allen Änderungen, sobald sie erfolgt sind, betroffen wäre.

Change Management ist insbesondere für ETL-Entwicklungsumgebungen wichtig, bei denen die Abhängigkeit der Datenelemente untereinander aus oberflächlicher Perspektive nur schwer zu erkennen ist. Change Management gestattet Entwicklern, ein bestimmtes Objekt (z. B. eine Tabellendefinition) zu überprüfen, mit der Auswirkung von Änderungen in einer geschützten Umgebung (einem Projektar-

beitsbereich) zu experimentieren und, nachdem deren Unbedenklichkeit festgestellt wurde, es in einen gemeinsamen Arbeitsbereich zurückzuleiten, damit andere Anwender es sehen können.

Sobald derartige Metadatenerfordernisse erwogen wurden, werden neue Upgrades möglich, die die Effektivität des ETL-Prozesses verbessern. Diese Upgrades werden durch zwei neue Services möglich: Multilevel Planning and Administration und Data Quality Integration.

## **Multiusser Planning**

Ein gewisses Mass an Vorüberlegungen ist erforderlich, wenn mehrere Entwickler, Administratoren oder Endanwender in ein Projekt miteinbezogen werden sollen. Beispielsweise erfordert die Miteinbeziehung mehrerer Entwickler in den Data-Warehouse-Entwurfs- und Implementierungsprozess die Einführung einer gemeinsamen Vorgehensweise bei der Speicherung von Dokumentations- und Implementierungsdaten. Dies ist wichtig, um zu gewährleisten, dass die Arbeit verschiedener Warehouse-Entwickler in eine konsistente und wartbare Struktur integriert werden kann. Zusätzlich müssen zur Handhabung von Multilevel Migration Kernmetadaten für multiple, verteilte Standorte verfügbar sein

und ihre Wiederherstellung durch Sicherungskopien gewährleistet sein, um gegen Hardwareversagen abgesichert zu sein.

### Multiuser Administration

Die grosse Anzahl an Personen, die typischerweise an einem unternehmensweiten Data-Warehousing-Projekt beteiligt sind, macht es notwendig, dass die verwendeten Werkzeuge auf die Rolle dieser Einzelpersonen zugeschnitten sind und den Administratoren dabei helfen,

- ▶ Projektarbeitsbereiche für jeden ETL-Entwickler zu erstellen
- ▶ Gruppenrechte für alle ETL-Entwickler zu vergeben
- ▶ allgemeine Informationen je nach Bedarf gemeinsam zu nutzen
- ▶ zu gewährleisten, dass Aktualisierungen unter Einsatz eines Audit-Trails erfolgen
- ▶ Check-out-/Check-in-Verfahren zur Vermeidung von Überschreibungen einzusetzen.

Der Wert eines gemeinschaftlich verwendeten Metadaten-Repositories wird deutlich, wenn die Informationen in einem Management-Cockpit sichtbar werden. Ein Administrator muss komplexe Definitionen für Daten- und Anwendungs-Server nur einmal festlegen und kann dann die Informationen mit der Anwendergemeinschaft gemeinsam benutzen.

Mit dem Vorhandensein einer Management Console lässt sich dies problemlos bewältigen.

### Data Quality Integration

Keine Abhandlung über Data Warehousing wäre ohne das Thema Datenqualität komplett. Wann immer Daten aus einer Vielzahl von Quellen zusammengestellt werden, wird die Minimierung der Auswirkungen

mangelhafter Daten zu einer Hauptaufgabe. Ein Data Quality Server kann die Aufgabe der Bereinigung fehlerhafter Daten und das Abgleichen von Daten aus unterschiedlichen Quellen mit allgemeingültigen Definitionen übernehmen. Dazu gehört u. a.

- ▶ die Erstellung von Match-Codes für Transformation, Reporting und Analyse zur Erzeugung und Anwendung von Schemata die innerhalb eines ETL-Transformations-Prozesses verwendet werden können
- ▶ die Bestimmung von Gross- oder Kleinbuchstabenstandardisierung
- ▶ die Nutzung anderer Funktionen zur Bestimmung des Geschlechts oder des Orts

### Der ETL-Entwurfsprozess

Der eigentliche ETL-Entwurfsprozess beginnt sobald die administrativen Funktionen eingerichtet und betriebsbereit sind. Von allen Data-Warehousing-Entwicklungsaktivitäten ist der ETL-Prozess gewöhnlich der zeitaufwendigste. Er ist schwierig zu entwickeln und zu betreiben. Die anfänglichen Aufgaben in einem ETL-Werkzeug bestehen in der Definition der zu extrahierenden Quellen und der zu ladenden Zieltabellen. Beispielsweise kann ein ETL-Entwickler eine ganz einfache Aufgabe haben: die grundlegenden Warehouse- und Transaktions-Daten in eine wöchentliche Übersichtstabelle umwandeln. Typischerweise würde dies innerhalb eines Projektbereichs erfolgen, um das Risiko und den Umfang der Arbeit einzuschränken. Anschliessend würde es in einen für andere Personen einzusetzenden gemeinsamen Bereich eingestellt werden. Des weiteren würden Gruppierung festgelegt werden, um die Teile zusammenzufügen und ein Verwechseln mit Komponenten anderer Projekte zu vermeiden. Diese Aufgaben können mit heutigen ETL-Werkzeugen problemlos implemen-

tiert und einfacher denn je ausgeführt werden und vereinfachen den Datenmanagementprozess so entscheidend.

### Erstellung eines Common Warehouse Metamodels

Die neuesten Data-Warehousing-Produkte unterstützen grösstenteils den offenen Common Warehouse Metamodel (CWM) Standard. Dieser gestattet die gemeinsame Nutzung von Metadateninformationen über Anwendungen hinweg. In einem typischen Grossunternehmen steuert ein Datenarchitekt das Unternehmens- oder Abteilungsdatenmodell. Dieses Datenmodell beschreibt, wie verschiedene Teile zueinander in Beziehung stehen, und sorgt für die "Single Version of Truth". Mit Datenmodellierungswerkzeugen wie z. B. ErWin oder Rational Rose werden Datenmodelle für den Einsatz definiert. Diese können direkt in viele ETL-Werkzeuge importiert werden. Das Ergebnis eines CWM-Imports sind alle Tabellendefinitionen, aus denen sich das Modell zusammensetzt. Auf diese Weise können eine beliebige Zahl von Zieldefinitionen von der Datenmodell-Entwurfsumgebung in die ETL-Entwicklungsumgebung übernommen werden.

### Schlussfolgerung

Durch neue Entwicklungen im Metadatenprozess und der Verfügbarkeit neuer Services wie Data Quality und Multiuser Planning wird Data Warehousing weiterhin neue Funktionalitäten bieten, die die Herausforderungen des sich ständig wandelnden Business Intelligence-Softwaremarkts adressieren. Wenn Führungskräfte in der Geschäftswelt grösseres Vertrauen in die Qualität ihrer Daten setzen, dann können sie bessere, sachlich begründete Entscheidungen treffen. IT-Manager können sich darauf verlassen, dass die ETL-Umgebung innerhalb ihrer Infrastruktur stabil und sicher ist. Und End-to-End-Lösungen, die hohe Datenqualität, Datenverteilung und leistungsstarke Analytik beinhalten, helfen Unternehmen für neue Entwicklungen jederzeit gerüstet zu sein.