DWH Automatisierung mit Data Vault 2.0

Andre Dörr
Trevisto AG
Nürnberg

Schlüsselworte
Architektur, DWH, Data Vault

Einleitung
Wenn man die Entwicklung von ETL / ELT Prozessen für eine klassische DWH Architektur (Kimball, Inmon) mit den Fertigungsprozessen der Automobilindustrie vergleicht, bewegt man sich hier noch in den 80er Jahren. Es herrscht so gut wie keine Automatisierung und Standardisierung. Änderungen und Erweiterungen erfordern zum Teil große manuelle Aufwände. Von einer schnellen Anpassungsfähigkeit kann man bei solch einem System nicht reden.
Um die DWH Entwicklung zu automatisieren und somit in die nächste Generation zu führen, wird ein neues DWH Konzept benötigt, welches genau diese Anforderungen nach Automatisierung und Standardisierung bietet. Nur so ist es möglich, agile Entwicklungsmethoden zu nutzen und die immer schnelleren Änderungen und Erweiterungen zu bewältigen.
Data Vault 2.0 bietet dieses Potential. Die Version 2.0 erweitert die Flexibilität des Datenmodells aus Version 1.0 mit einer passenden Architektur. Im Gegensatz zu den klassischen Ansätzen ist damit eine Automatisierung der Entwicklung sehr einfach möglich. Dies kann mit dem richtigen Tool-Set zu einer drastischen Verkürzung der Entwicklungszeiten führen.

Was versteht man unter DWH Automatisierung?

Unter DWH Automatisierung versteht man im Allgemeinen die automatisierte Generierung von Datenstrukturen und ETL / ELT Prozessen. Diese Automatisierung gewinnt immer mehr an Bedeutung bei der Entwicklung und Weiterentwicklung eines DWH Systems. Anforderungen der Fachbereiche müssen immer schneller umgesetzt werden, was bereits dazu geführt hat, dass agile Vorgehensmodelle eingesetzt werden. Jedoch sollte auch die Architektur eines DWH Systems eine agile Entwicklung unterstützen. Die Architektur muss einen gewissen Grad an Automatisierbarkeit bieten, so dass Anpassungen auch innerhalb der kurzen Entwicklungszyklen umgesetzt werden können. Data Vault in der Version 2.0 bietet dafür 2 Ansatzpunkte: Zum einen, ist das Datenmodell so standardisiert, dass eine vollkommen automatisierte Generierung möglich ist. Zum zweiten, wird dies noch durch die Referenz-Architektur und die Aufteilung in 4 verschiedene Schichten unterstützt.

Die Farben der Datenmodellierung

Das Datenmodell von Data Vault 2.0 vereint die Vorteile des 3NF Datenmodells und des dimensionalen Datenmodells. Es wurde speziell für die historische Speicherung von Daten innerhalb eines DWHs entwickelt. Vergleicht man die Datenhaltung innerhalb eines Data Vault Modells mit der eines 3NF oder dimensionalen Modells, wird schnell deutlich, wie Data Vault standardisiert ist, und wie dies eine automatisierte Generierung unterstützt. Alle Daten innerhalb eines DWHs können in drei verschiedene Klassen unterteilt werden. Business Keys identifizieren die einzelnen Geschäftsobjekte eindeutig. Assoziationen beschreiben die Beziehung zwischen den einzelnen Geschäftsobjekten. Alle weiteren Daten sind beschreibende Attribute für die Geschäftsobjekte oder Beziehungen. Vergibt man für alle drei Klassen verschiedene Farben, kann man die Verteilung dieser Daten innerhalb der verschiedenen Modellierungsvarianten sehr gut darstellen.

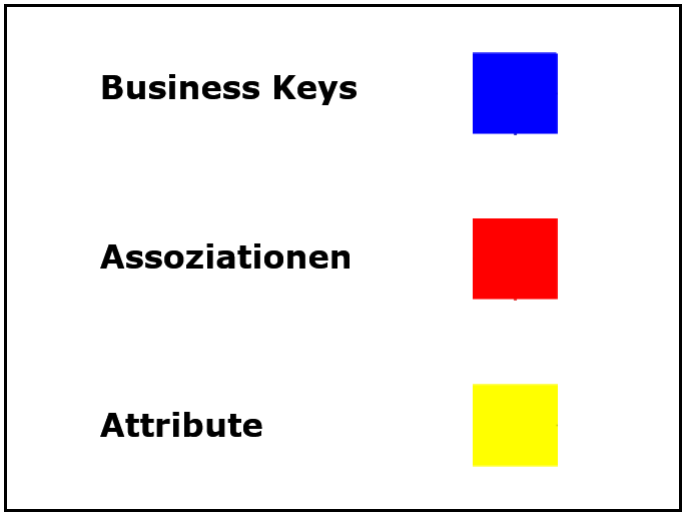

Abbildung 1: Die Farben der Datenmodellierung

Ein 3NF Datenmodell kennt grundsätzlich nur eine Art von Tabellen. Jede Tabelle innerhalb eines 3NF Modells beinhaltet somit auch alle drei Arten von Daten. Eine 3NF-Tabelle besitzt einen Primary- / Business Key für die eindeutige Identifizierung eines Objekts. Die Foreign Keys einer Tabelle entsprechen den Beziehungen zwischen den verschiedenen Objekten. Alle weiteren Spalten sind die beschreibenden Attribute eines Objekts. Stellt man ein 3NF Modell mit den verschiedenen Farben dar, ergibt sich somit auch ein sehr gemischtes Bild.

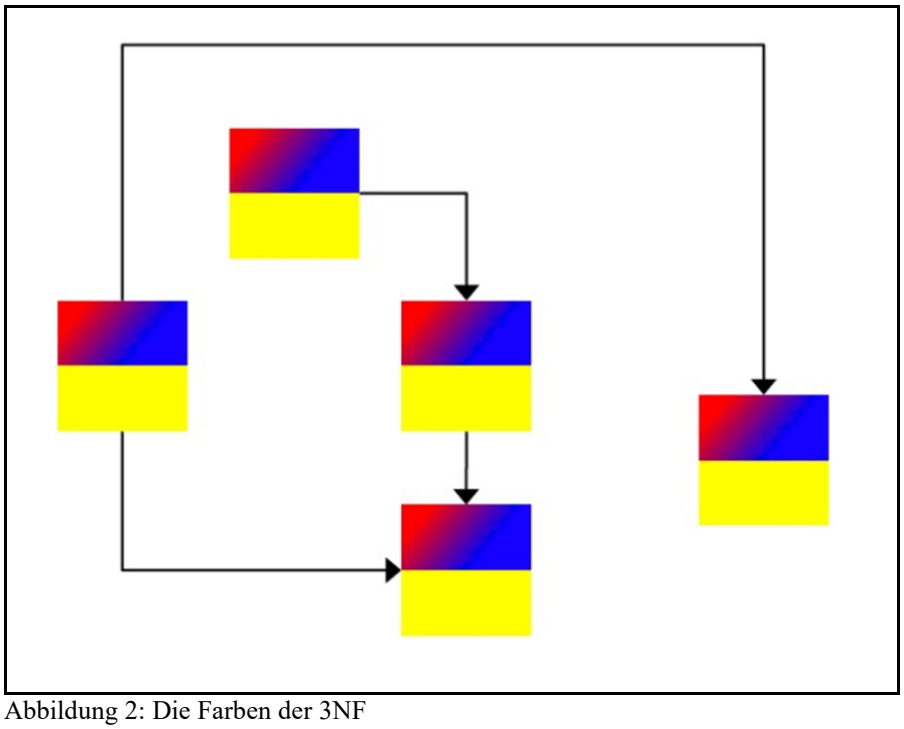
Abbildung 2: Die Farben der 3NF

Ein dimensionales Datenmodell besteht aus zwei verschiedenen Arten von Tabellen: Fakten-Tabellen und Dimensionstabellen. Dimensionstabellen bestehen zum größten Teil aus beschreibenden Attributen. Die restlichen Bestandteile sind Business Keys für die eindeutige Identifizierung der einzelnen Hierarchie-Ebenen. Die Faktentabelle beinhaltet alle drei Arten von Daten. Die Foreign Keys auf die verschiedenen Dimensionen repräsentieren die Beziehungen zwischen den verschiedenen Objekten. Die Fakten selbst entsprechen auch beschreibenden Attributen. Eine Faktentabelle kann zusätzlich noch eindeutig identifizierende Attribute beinhalten. Die farbliche Verteilung der Daten innerhalb eines dimensionalen Datenmodells zeigt bereits etwas mehr Strukturierung als bei einem 3NF Modell.

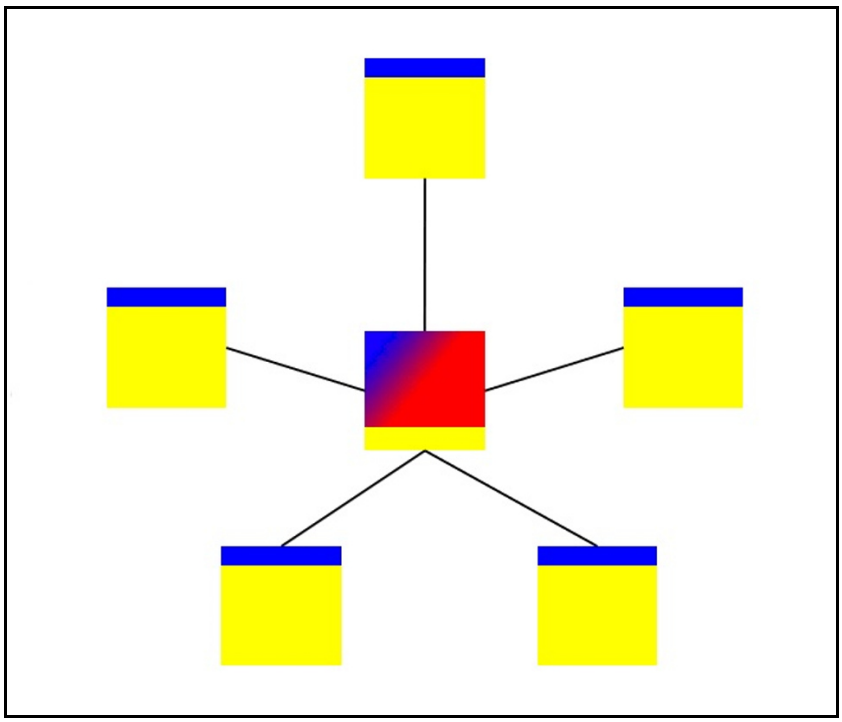

Abbildung 3: Die Farben des Star Schemas

Ein Data Vault Datenmodell besteht aus drei verschiedenen Tabellen-Typen: Hubs, Satelliten und Links. Die Grundidee der Data Vault Modellierung besteht aus der Trennung von identifizierenden und beschreibenden Attributen. Hub Tabellen speichern die Business Key Information für die einzelnen Geschäftsobjekte. Link Tabellen beinhalten die Beziehungsinformationen zwischen den verschiedenen Geschäftsobjekten. Alle beschreibenden Informationen werden in Satelliten gespeichert. Man unterscheidet zwischen Hub-Satelliten und Link-Satelliten. Hub-Satelliten beinhalten die beschreibenden Informationen zu Geschäftsobjekten. In Link-Satelliten werden beschreibende Informationen zu den Beziehungen gespeichert. Durch diese Aufteilung der Daten ergibt sich ein sehr strukturiertes Bild.

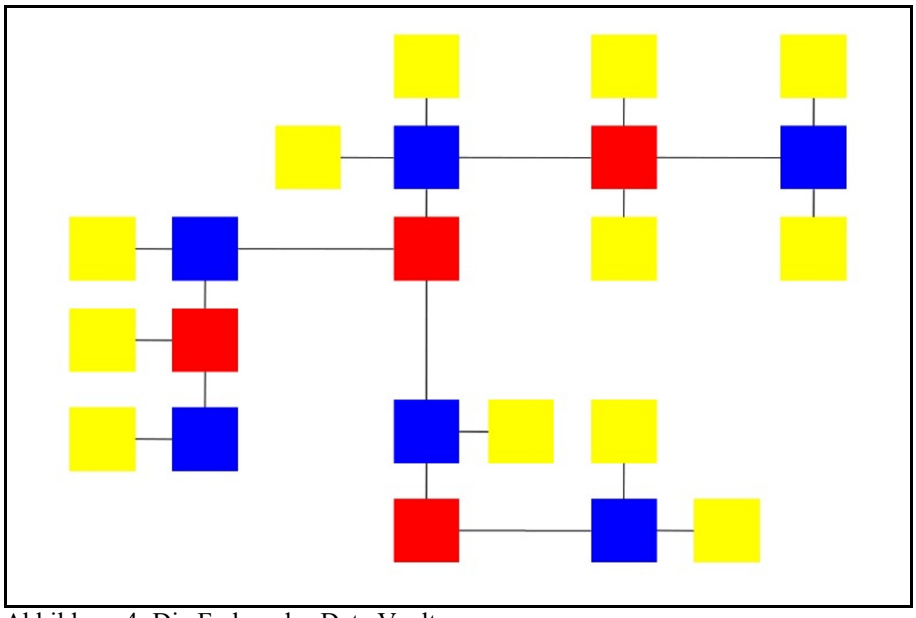

Abbildung 4: Die Farben des Data Vaults

Im Vergleich zur 3NF und dimensionalen Datenmodellierung beinhaltet jede Tabelle in einem Data Vault Modell nur eine Art von Daten. Dadurch ergibt sich eine Standardisierung der Datenstrukturen. Jede Hub-Tabelle, jede Satelliten-Tabelle und jede Link-Tabelle hat den gleichen Grundaufbau. Dies führt auch zu einer Standardisierung der Ladeprozesse. Für Tabellen mit der gleichen Datenstruktur können auch die gleichen Ladestrukturen verwendet werden. Man spricht hierbei von „Pattern Based Loading“. Das Datenmodell basiert auf immer wiederkehrenden Mustern. Diese Muster bilden die Voraussetzung für eine einfache Generierbarkeit und Automatisierbarkeit.

Trennung von Soft- und Hard-Rules

Data Vault unterstützt jedoch nicht nur durch die Datenmodellierungsregeln eine DWH Automatisierung. Data Vault 1.0 definierte die Regeln für das Datenmodell. Data Vault 2.0 ergänzt dieses um eine Referenz-Architektur und ein agiles Vorgehensmodell.
Ein klassisches DWH nach Kimball oder Inmon besteht in der Regel aus 3 Schichten. Einem Staging-Layer, dem Core-DWH und dem Mart-Layer. Man spricht hierbei gerne von dem Single-Point-of-Truth. Dieser Ansatz spiegelt jedoch nicht die Anforderungen an eine agile Entwicklung wieder. Anpassungen erfordern zum Teil einen hohen manuellen Aufwand.
Um dieses Problem zu umgehen, führt die Data Vault Referenz Architektur eine strikte Trennung von Hard-Rules und Soft-Rules ein. Hard-Rules beinhalten alle Datenverarbeitungsregeln, welche nicht den Inhalt der Daten verändern (z.B.: Datentypkonvertierungen). Diese werden auf dem Weg in den historisierten Raw Data Vault Layer angewendet. Soft-Rules entsprechen den eigentlichen Geschäftsregeln. Diese werden jedoch erst nach der persistenten Speicherung der historischen Daten angewendet. Durch diese Trennung können die Geschäftsregeln schnell an neue Anforderungen angepasst werden, ohne dass Änderungen an den persistent historisch gespeicherten Daten vorgenommen werden müssen.

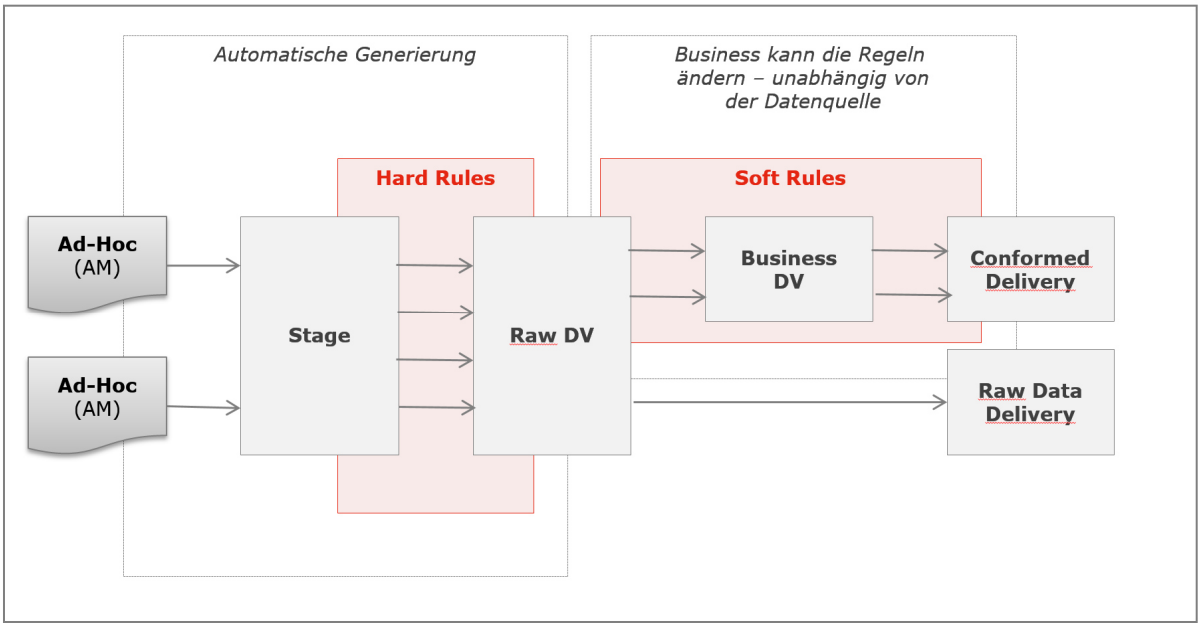

Abbildung 5: Data Vault 2.0 Referenz-Architektur

Die Data Vault Architektur unterstützt durch diesen Aufbau jedoch nicht nur den Einsatz von agilen Entwicklungsmethoden, es vereinfacht auch die automatisierte Generierung von Datenstrukturen und Datenverarbeitungsprozessen. Geschäftsregeln (Soft-Rules) können eine beliebige Komplexität annehmen. Da diese jedoch innerhalb der Architektur nach hinten verlagert wurden, können alle Strukturen und Prozesse bis zur Rohdatenschicht sehr einfach automatisiert generiert werden.

Vorteile der DWH Automatisierung

Eine Automatisierung der DWH Entwicklung bietet viele Vorteile. Durch sinkende Entwicklungsaufwände kann schneller auf sich ändernde Anforderungen reagiert werden. Erst so ist ein effizienter Einsatz von agilen Entwicklungsmethoden möglich. Die Standardisierung trägt dazu bei, das Kosten und Aufwände für die Konzeption bis zur Wartung reduziert werden.
Eine DWH Automatisierung kann dabei mit Tool-Unterstützung als auch manuell umgesetzt werden. Aktuell wächst du Anzahl an Anbietern für Automatisierungstools immer mehr. Folgende Tools haben sich bereits bewährt:

- Quipu
- WhereScape
- BI Ready
- Data Vault Builder

Alle Tools bieten einen unterschiedlichen Ansatz und Funktionsumfang. Quipu beschränkt sich z.B. auf die Generierung der Rohdaten-Schicht. WhereScape hingegen gehört zu den Tools, mit denen auch die Soft-Rules für die Data Vault Architektur umgesetzt werden können. Für die Generierung der Datenstrukturen und ELT-Prozessen werden in der Regel individualisierbare Templates eingesetzt, so dass der erzeugte Code auf die Anforderungen der verschiedenen Datenbanken angepasst werden kann.

Kontaktadresse:
Andre Dörr
Trevisto AG
Nunnenbeckstraße 6/8
D-90489 Nürnberg

Telefon:	+49 (0) 911-430 839 00
Fax:	+49 (0) 911-430 839 01
E-Mail	amdre.doerr@trevisto.de.de
Internet:	www.trevisto.de