OPITZ CONSULTING

Menschen. Innovationen. Lösungen.

Fact Sheet

Data Vault – die Alternative für agile Data-Warehouse-Projekte

Agile DWH-Projekte erfordern ein flexibles und dennoch stabiles Datenmodell, genau hier punktet die Modellierungsmethode Data Vault gegenüber 3NF

Die Modellierung mit Data Vault stellt dank ihrer hohen Flexibilität bei Erweiterungen eine ideale Grundlage für agile Projekte dar. Zudem kann aktuelle Hardware durch die klaren Abhängigkeiten bei der Datenbewirtschaftung mithilfe von Parallelisierung ideal ausgelastet werden. So lassen sich die Laufzeiten gering halten. Doch auch wenn bei der Erzeugung eines Data Vault Modells nur wenige Regeln zu beachten sind, bieten sich in der Praxis häufig für eine Lösung mehrere Wege der Modellierung an. Unsere erfahrenen Berater unterstützen Sie bei der Wahl des für Sie besten Wegs!

In den letzten Jahren hat sich mit "Data Vault" eine neue DWH-Modellierungsmethode profiliert, die für Integrationsschichten eine Alternative zu klassischen Modellierungsmethoden darstellt. Data Vault bietet ein einfaches Grundmodell mit wenigen Basiskonzepten, das es erlaubt, massiv parallelisierbare Ladeprozesse zu nutzen. Die strukturelle Entkopplung der Daten ermöglicht eine agile und einfache Erweiterbarkeit des Datenmodells und des Data Warehouse.

Das Besondere an Data Vault

Im Gegensatz zum Klassiker der Modellierung, der dritten Normalform (3NF), werden bei der Modellierung mit Data Vault alle zu einem Objekt gehörenden Informationen in drei Kategorien unterteilt und strikt getrennt voneinander abgelegt. In die **erste Kategorie** gehören Informationen, die einem Objekt seine Identität geben. Bei Kunden kann dies z. B. die Kundennummer sein oder bei Artikeln die Artikelnummer. Entscheidend ist, dass diese Informationen ein Objekt eindeutig identifizieren. Diese Schlüsselinformationen werden in sogenannten Hubs abgelegt, die in den Abbildungen auf der rechten Seite blau hinterlegt sind. Attribute, die ein Objekt grundlegend beschreiben, fallen in die **zweite Kategorie**. Hierzu zählen Merkmale wie der Kundenname, die Artikelbezeichnung oder die Artikelfarbe. Informationen dieser Art werden in „Satelliten" abgelegt und sind in den Abbildungen gelb markiert. Die **dritte Kategorie** bilden Informationen, die die Beziehungen zwischen zwei oder mehr Objekten beschreiben. Die Zuordnung eines Kunden zu einem Kundentyp oder einem Händler ist eine solche Beziehung. Informationen dieser Art werden in „Links" geschrieben und sind in den Abbildungen rot hinterlegt.

Die Abbildungen veranschaulichen das Vorgehen am Beispiel einer Kundentabelle. Diese ist einmal in der dritten Normalform und einmal als Data Vault modelliert. Jede Tabelle enthält nur Attribute aus einer der drei genannten Kategorien.

Datenmodelle im Vergleich

Modellierung in dritter Normalform

KUNDE		
PK	KUNDE_ID	Schlüssel
UK	KUNDENNR	
	TITEL	Beschreibungen
	VORNAME	
	NACHNAME	
	ADRESSE	
FK	KUNDENTYP_ID	Beziehungen
FK	HÄNDLER_ID	

Modellierung mit Data Vault

H_KUNDE	
PK	H_KUNDE_ID
UK	KUNDENNR_NK
	ZEITSTEMPEL
	QUELLE

L_KUND_KUTY	
PK	KUND_KUTY_ID
FK	H_KUNDE_ID
FK	H_KUNDENTYP_ID
	ZEITSTEMPEL
	QUELLE

S_KUNDE	
PK	H_KUNDE_ID
PK	ZEITSTEMPEL
	QUELLE
	TITEL
	VORNAME
	NACHNAME
	ADRESSE

L_KUND_HÄND	
PK	KUND_HÄND_ID
FK	H_KUNDE_ID
FK	H_HÄNDLER_ID
	ZEITSTEMPEL
	QUELLE

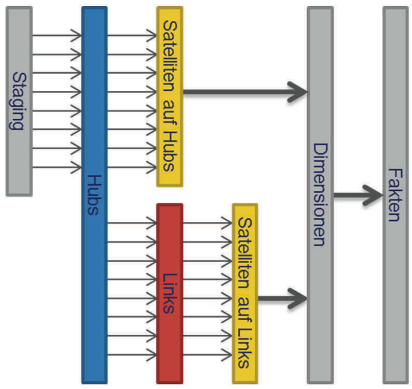
Klare Trennung von Identitäten, Eigenschaften und Beziehungen

Die Tabelle in der dritten Normalform ist bereits redundanzfrei und wird in der Modellierung mit Data Vault durch vier Tabellen ersetzt. Werden die Abhängigkeiten der Tabelle KUNDE betrachtet, müssen erst die referenzierten Tabellen KUNDENTYP und HÄNDLER aktualisiert sein, bevor die Beladung von Tabelle KUNDE erfolgen kann.

Genau diese erhöhte Komplexität bildet die Basis für die wesentlichen Vorteile von Data Vault. Denn was bei kleinen Modellen noch überschaubar ist, wird in komplexen Modellen schnell unübersichtlich und damit fehleranfällig. Die Verteilung auf mehrere Tabellen löst dieses Problem.

Das Data Vault Modell ist immer auf die gleiche Weise zu laden und die Beladung ist dazu noch sehr einfach zu parallelisieren, da es keine Abhängigkeiten zwischen Objekten des jeweils gleichen Kategorie gibt. Die Abbildung zeigt wie sich die Abhängigkeiten bei Data Vault gestalten:

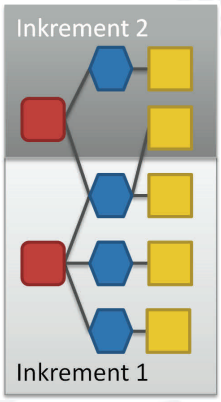
Abhängigkeiten bei Data Vault



Inkrementell erweiterbar

Neben der guten Parallelisierbarkeit der Ladeprozesse und der entkoppelten Abhängigkeiten bietet diese strikte Trennung weitere Vorteile gegenüber der herkömmlichen Modellierung. Muss ein solches Datenmodell erweitert werden, weil z. B. Daten aus einem weiteren Quellsystem integriert werden sollen oder sich die Anforderungen ändern, sind in der Regel keine bestehenden Tabellen zu modifizieren.

Stattdessen werden bei jedem weiteren Inkrement nur Tabellen hinzugefügt. Dies reduziert die Ausfallzeit während einer Auslieferung oder vermeidet sie gar vollständig.



Liefert das neue System beispielsweise weitere Informationen zum KUNDEN, wird hierfür einfach ein weiterer Satellit angelegt, der den gleichen Hub referenziert wie der bereits existierende Satellit. Eine weitere Anpassung des Datenmodells ist nicht erforderlich.

Auch der Einfluss auf die Beladungsprozesse ist minimal. Lediglich die Beladung des Kunden-Hubs muss um eine Quelle ergänzt werden. Da Hubs immer auf dieselbe Weise geladen werden, werden hier je nach ETL-Werkzeug konfigurierbare Templates verwendet, so dass diese Anpassung im besten Fall nur eine Änderung der Konfiguration erfordert.

Transparenz und zeitliche Nachvollziehbarkeit

Bei der Modellierung mit Data Vault stehen noch zwei weitere grundlegende Prinzipien im Vordergrund. Bei jedem Datensatz muss erstens erkennbar sein, wann er geladen wurde und zweitens woher er stammt. Hubs und Links sind generell zeitlich unveränderlich. Es kommen also nur neue Schlüssel hinzu.

Bei Satelliten verhält sich dies anders. Hier ist durch Zeitstempel immer erkennbar, seit wann diese Information gültig ist bzw. wann sie vorlag. In manchen Implementierungen wird an Stelle des Zeitstempels ein klassisches Gültigkeitsintervall verwendet. Dieses ist aus Abfragesicht sicherlich komfortabel, erfordert aber Aktualisierungen der Satellitentabelle, was im Hinblick auf Big-Data-Architekturen häufig ungünstig ist.

Sie möchten mehr über das Thema oder zu unserer Expertise im Bereich Data Vault Modellierung erfahren? Sie wünschen eine individuelle Beratung? Sprechen Sie mich an:

Marcel Aretz
Senior Consultant
Telefon: +49 40 741122-0
Telefax: +49 40 741122-4300
E-Mail: marcel.aretz@opitz-consulting.com

Mehr zu unserem Leistungsbereich Business Information Management:
www.opitz-consulting.com/business_information_management

Folgen Sie uns

www.opitz-consulting.com/newsroom