Data Vault – der Schlüssel zum Wachstum

Architekturen in der Praxis

Inmon oder Kimball? Das ist seit jeher die entscheidende Frage für BI-Architekten, wenn es um den Aufbau eines Data Warehouse (DWH) geht. Über die Vor- und Nachteile beider Ansätze entbrannte im Verlauf der Jahre ein wahrer Lagerkampf, aus dem bis heute kein eindeutiger Sieger hervorgegangen ist. Mit der Data-Vault-Methode etabliert sich gegenwärtig eine Vorgehensweise, die zum Schulterschluss der großen Vorbilder beiträgt und zugleich das DWH für die Herausforderungen der Zukunft rüstet.

Fakt ist: Sowohl der normalisierte Ansatz von Bill Inmon als auch die dimensionale Vorgehensweise nach Ralph Kimball sind in theoretischer Hinsicht sinnvoll und praktikabel. Wenn sie scheitern, dann an der allgegenwärtigen Hektik des Unternehmensalltags – sprich: an schnell wachsenden und wechselnden Anforderungen, mangelnder Abstimmung zwischen den Abteilungen, einem konstanten Zeitdruck sowie begrenzten Budgets. Den daraus resultierenden Wildwuchs in der DWH-Architektur kann weder das eine noch das andere Modell kompensieren. Langfristig leiden darunter sowohl die Betriebsstabilität als auch die Konsistenz der Daten.

Können also die großen Vorbilder den Herausforderungen einer modernen Geschäftswelt und der immer rasanteren technischen Entwicklung nicht mehr standhalten? Betrachtet man den weitaus jüngeren Data-Vault-Ansatz von Dan Linstedt, liegt diese Vermutung zunächst nahe. Schließlich verspricht er eine hohe Flexibilität bei Erweiterungen und somit ein homogenes Wachstum. Aber handelt es sich tatsächlich um eine vollkommen neue und eigenständige Methodik? Lässt sich die erfolgreiche Integration eines DWH bei den vielfältigen Ansprüchen und divergierenden Infrastrukturen in Unternehmen überhaupt noch mit einem einzigen Vorgehensmodell gewährleisten? Und welche Architektur ist im konkreten Fall tatsächlich erfolgversprechend? Ein Blick in die Anwendungspraxis der verschiedenen Vorgehensweisen gibt Aufschluss.

Inmon: solide, aber unflexibel

Die klassische Inmon-Architektur ist vor allem in größeren deutschen Unternehmen weit verbreitet. Sie gilt als „solide Wertarbeit“, was vor allem darauf zurückzuführen ist, dass unmittelbar sämtliche Unternehmensbereiche in einem Datenmodell integriert werden. Es wird also eine umfassende und vermeintlich stabile Basis geschaffen. Und tatsächlich weist diese Vorgehensweise gewisse Vorteile auf: Anstelle von Aggregaten liegen die Daten im kleinstmöglichen Grain (Quäntchen) vor. Historische Daten sind jederzeit abrufbar. Neue Data Marts, über die der Fachanwender auf die Daten zugreift, lassen sich sehr einfach bereitstellen. Nicht zuletzt stellt der Ansatz die „Single Version of Truth“ sicher – also eine einheitliche Definition der im Unternehmen kursierenden Kennzahlen.

Wie aber die Praxis zeigt, hat die Solidität auch ihren Preis. Und das im wahrsten Sinne des Wortes: Bereits die Modellierung der Basis unter Berücksichtigung aller Fachabteilungen ist äußerst komplex und kostenintensiv. Bis das System endlich produktiv ist, werden Entscheider und Nutzer vor eine lange Geduldsprobe gestellt. Ähnlich verhält es sich mit der Weiterentwicklung des Datenmodells. Im Zuge der hohen Komplexität ist eine iterative beziehungsweise agile Vorgehensweise weitestgehend ausgeschlossen. Änderungen führen sowohl hinsichtlich der Beladung als auch der Data Marts zu hohen Folgekosten. Gleichzeitig besteht die Gefahr, dass die Fachbereiche nicht entsprechend ihren Bedürfnissen bedient werden. Denn: Im Regelfall ist der Organisationsaufwand so groß, dass allein für das Core Warehouse ein eigenes Team bereitgestellt werden muss. Hierbei handelt es sich erfahrungsgemäß um technische Spezialisten, die den Fokus auf die Datenintegration und weniger auf fachliche Fragestellungen legen. Das Problem potenziert sich, wenn derlei Tätigkeiten nach draußen verlagert werden.

So oder so bleibt es schwierig, das Core-Modell und die Data Marts dauerhaft konsistent zu halten. Meist gilt es, eine vierstellige Zahl von Tabellen zu beherrschen, wobei ein Architekt bereits mit etwa 200 Stück voll ausgelastet ist. Die erforderliche Manpower wie auch ein entsprechendes Budget sind nur selten gegeben. Daher werden gerne „Abkürzungen“ umgesetzt, durch die das System zunehmend verwuchert. Ebenso legen Unternehmen oftmals Dutzende von Data Marts an, für die eine entsprechende Pflege kaum noch zu leisten ist. Neben Inkonsistenzen ist ein zunehmender Anforderungsstau das Resultat.

Insofern ist auch die von Inmon-Verfechtern gepriesene, mutmaßlich höhere Datenqualität sehr differenziert zu betrachten. Grundsätzlich zeigt die Erfahrung, dass der Einfluss des Datenmodells auf die Qualität überbewertet wird. So lassen sich beispielsweise mit Constraints und Foreign Keys nur ein Teil der zahlreichen Datenregeln sicherstellen. Viele weitere Regeln können nur bei der Beladung unterstützend wirken. Nicht zuletzt ist festzuhalten, dass man bei einem hochgradig normalisierten Modell viel zu wenig von den heutzutage möglichen Datendurchsätzen und -volumen profitieren kann.

Kimball: Schnelle Umsetzung mit Nutzwert

Inmon legt also den Schwerpunkt auf die korrekte und vollständige Speicherung der gesamten Unternehmensdaten. Der unmittelbare Nutzwert für den Fachanwender wird nachgeordnet behandelt. Ganz anders Ralph Kimball: Für ihn steht die einfache Bereitstellung stets im Vordergrund. Dabei orientiert er sich strikt an Geschäftsprozessen und aktuel-

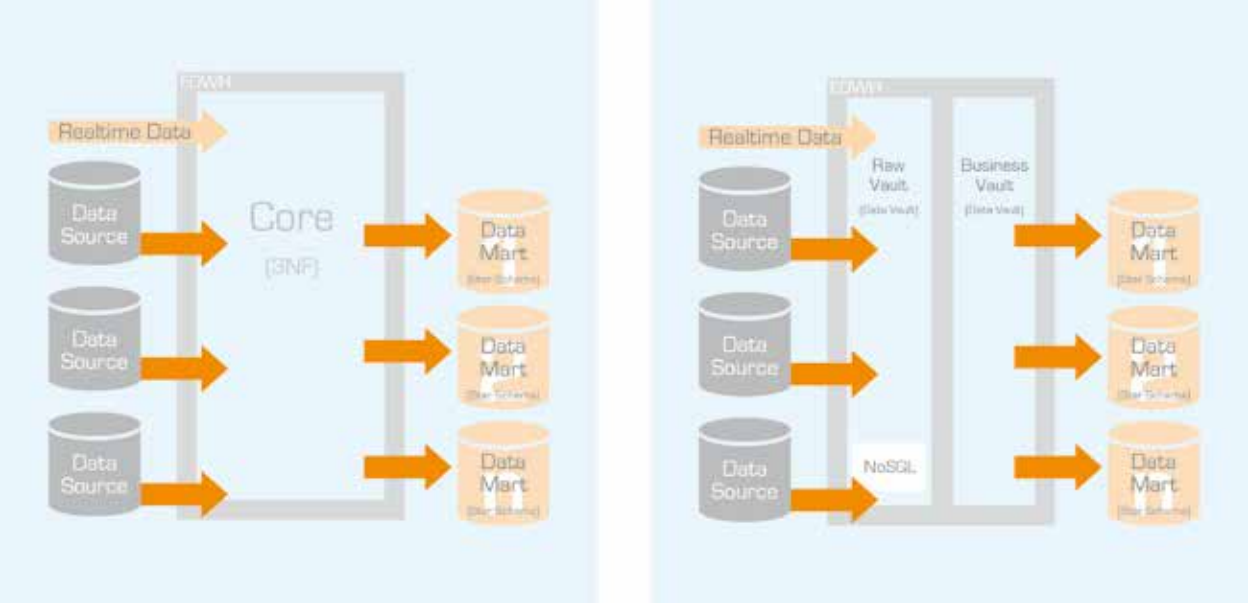

Abb. 1: Inmon in der klassischen Ausführung (links) und ergänzt durch Data Vault (rechts): Linstedts Ansatz sorgt für Flexibilität und stabiles Wachstum im Core DWH. Unabhängig davon hat sich eine Modellierung der Data Mart nach Kimballs Star-Schema als sinnvoll erwiesen. (Quelle: ORAYLIS GmbH)

len Anforderungen. Ein erster Prototyp mit den wichtigsten Basisfunktionalitäten kann frühzeitig in Betrieb genommen werden. Weitere Funktionen werden iterativ beziehungsweise nach und nach modelliert. Der Aufbau ist für den Nutzer transparent und leicht verständlich. Aufgrund der schnellen Lieferung ist zudem eine hohe Akzeptanz gewährleistet.

Ein entscheidender Vorteil dieser Vorgehensweise ist, dass agile Entwicklungsprozesse umfassend unterstützt werden. Wie zahllose Projekte der Vergangenheit gezeigt haben, lässt sich Business Intelligence mit sequenziellen IT-Methoden nur schwerlich aufbauen. Die komplexen Anforderungen eines BI-Umfeldes können niemals in aller Vollständigkeit vorhergesehen und geplant werden. Dem setzt die Kimball-Architektur ein hohes Maß an Flexibilität entgegen. Das Modell kann in jede Richtung wachsen, ohne dass Änderungen in den Clientanwendungen erforderlich wären. Kurskorrekturen sind jederzeit möglich – vorausgesetzt gewisse Regeln werden eingehalten. Kimball selbst gibt leider keine klaren Empfehlungen, ob und wie die Daten im sogenannten Backroom gespeichert werden sollten. Hier sind selbstdefinierte Standards und Best Practices aus erfolgreichen Projekten erforderlich, um ein frühzeitiges Scheitern zu vermeiden. Ebenso essenziell ist ein gutes Team mit entsprechender Expertise.

Wer sich strikt an die Standards hält, der kann eine dimensionale Modellierung relativ einfach durchsetzen. Dabei besteht der „Frontroom“ im Idealfall aus nur einem großen Data Mart. In der Praxis kommt meist noch eine – wohlgemerkt überschaubare – Anzahl hinzu. Die Integration erfolgt durch konforme Dimensionen. Weitere Datenqualitätsregeln werden auf der ETL-Strecke sichergestellt. Auf diese Weise lassen sich dann auch große Projekte umsetzen, bei denen Kritikern die Kimball-Methode oftmals als ungeeignet erscheint. Selbstverständlich ist es eine Herausforderung, eine Vielfalt komplexer Sachverhalte in einem hochgradig integrierten Modell zu verknüpfen. Wer aber nach dem Prinzip „Jede vermiedene Tabelle ist eine gute Tabelle“ verfährt und zugleich eine gewisse Fehlertoleranz im Rahmen der Entwicklungsprozesse einräumt, der wird durch hervorragende Analyseeigenschaften und ausgezeichnete Ergebnisse belohnt. Darüber hinaus eignet sich die dimensionale Modellierung auch sehr gut für den Aufbau von Data Marts im Rahmen einer Inmon-Architektur.

Linstedt: Das Beste aus beiden Welten

Bei allen Vorteilen mangelt es dem Kimball-Ansatz vor allem an einem: klaren Prinzipien für die Modellierung des Backrooms. Derlei Vorgaben sind jedoch essenziell, um ein dimensionales Modell weiterzuentwickeln und typische Anforderungen – etwa die Abwärtskompatibilität oder die Vermeidung von Datenverlusten – zu erfüllen. Genau in diese Lücke stößt nunmehr Dan Linstedt mit der Data-Vault-Architektur. Sein Ansatz ist gezielt auf das stabile Wachstum eines DWH ausgelegt. Dabei eignet sich die Art der Modellierung gleichermaßen für den Backroom wie auch das Core Warehouse.

Linstedt bildet somit auch keine Konkurrenz zu Inmon oder Kimball. Vielmehr liefert er die Basis, um das Beste aus beiden Welten zu verbinden: Data Vault ist sowohl projekt- als auch ganzheitlich orientiert. Einerseits wird ein integriertes, unternehmensweites Modell ermöglicht, bei dem die Daten historisiert sowie im kleinstmöglichen Grain vorliegen beziehungsweise abrufbar sind. Andererseits unterstützt die Vorgehensweise agile Entwicklungsprozesse und eine schnelle Lieferung. Dies ist vor allem der Skalierbarkeit des Datenmodells zu verdanken, die durch eine konsequente Trennung zwischen identifizierenden und deskriptiven Attributen realisiert wird. Des Weiteren reduziert sich durch das iterative Vorgehen der Aufwand für Analyse und Modellierung. Entsprechend gering sind die Set-up-Kosten des Projekts. Im weiteren Verlauf zeigt sich Data Vault abso-

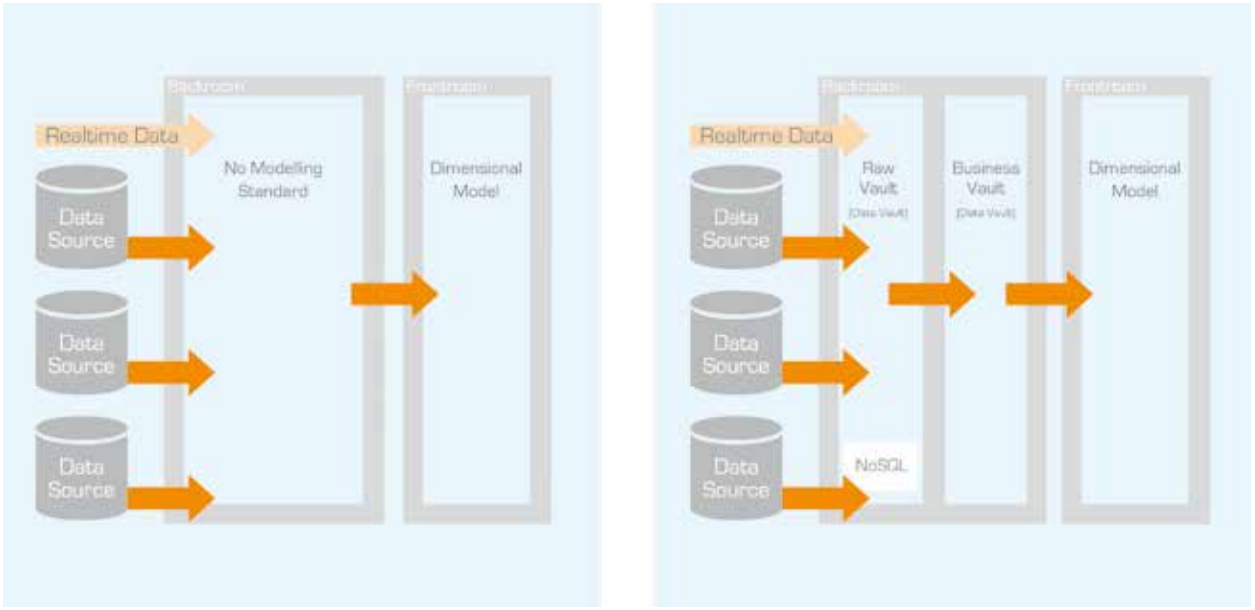

Abb. 2: Kimball ohne (links) und mit Data Vault (rechts): Hier schafft Linstedt klare Prinzipien bei der Modellierung des Backroom (Quelle: ORAYLIS GmbH)

lut flexibel hinsichtlich Veränderungen und Wachstum. Der Standard Data Vault 2.0 integriert sogar NoSQL-Lösungen wie Hadoop nahtlos, sodass man auch bestens für anstehende Big-Data-Szenarien gerüstet ist.

Zugleich wird durch einfach zu lernende, standardisierte Prozesse sowie die Reduzierungen von Abhängigkeiten in der Architektur die Skalierbarkeit von Teams optimal unterstützt, was sonst vor allem bei großen Projekten ein Problem darstellt. Ebenso lässt sich die Beladung dank hoher Standardisierung leicht generieren. In dem Fall beschränken sich die Tätigkeiten auf das Modellieren, Analysieren und gegebenenfalls Transformieren. Alles andere kann als Standard festgelegt werden. So wird die erforderliche Manpower gering gehalten. Es gibt kaum noch Fleißarbeit, sodass Off- oder Nearshoring entfallen können.

Ein vermeintlicher Nachteil des Data-Vault-Ansatzes ist sicherlich die mangelnde Abfrage- beziehungsweise Analyseorientierung. Es handelt sich explizit um ein datengetriebenes Modell, bei dem die Integration über Business Keys erfolgt. Die Interpretation der Daten ist also zunächst zweitrangig. Alles in allem weist auch diese Vorgehensweise einige Interpretationsspielräume auf, sodass es für die erfolgreiche Projektabwicklung eines erfahrenen Architekten bedarf. Dann aber ist das Modell einfach umzusetzen und für den Anwender sicher zu beherrschen.

Fazit

In der Unternehmensrealität ist heutzutage nichts so konstant wie die Veränderung. Und genau daran scheitern DWH-Projekte am häufigsten. Bei der Inmon-Architektur sind die erforderlichen Erweiterungen so aufwendig und teuer, dass auf der Suche nach günstigeren Lösungswegen oftmals das integrierte Datenmodell geopfert wird. Die Folge sind eine höhere Komplexität und ein gesteigerter Wartungsaufwand. Währenddessen besteht bei Kimball die Gefahr, dass jede Fachabteilung „ihr eigenes Süppchen kocht“ und somit die Konformität einzelner Dimensionen zunehmend aus dem Blick gerät. Hinzu kommt, dass ein DWH in der Praxis meist nicht „auf der grünen Wiese“ gebaut, sondern in eine bereits vorhandene Infrastruktur integriert werden muss.

Der Data-Vault-Ansatz erweist sich in verschiedensten Situationen als rettender Anker. Man kann ihn als eine Evolution von Inmon verstehen, die auch für komplexeste Systeme geeignet ist. Agilität, Geschwindigkeit und Wirtschaftlichkeit werden nachhaltig unterstützt. Nicht umsonst wird diese Vorgehensweise inzwischen sogar von Inmon selbst empfohlen. Kimball-Anhängern bietet Data Vault indes einen sehr guten und stabilen Unterbau für ihr weiteres Vorgehen. Linstedts Ansatz steht somit nicht im Wettbewerb zu den bewährten Architekturen. Er bildet vielmehr eine sinnvolle Ergänzung.

Bislang hat Data Vault in Deutschland noch nicht die gebührende Verbreitung gefunden. Entsprechend ist bei Entwicklern und Kunden noch eine gewisse Scheu vor dem Unbekannten zu beobachten. Nach ersten Erfahrungen stößt der Ansatz aber stets auf eine hohe Akzeptanz bei DWH-Entwicklern, Fachanwendern und Power-Usern. Es ist also noch einiges an Pionierarbeit zu leisten. Diese erweist sich aber in jeder Hinsicht als lohnenswert.

Daniel Piatkowski ist zertifizierter Data-Vault-Experte der ORAYLIS GmbH. **E-Mail: d.piatkowski@oraylis.de**
Thomas Strehlow verantwortet als Geschäftsführer die Delivery der ORAYLIS GmbH, Düsseldorf. **E-Mail: t.strehlow@oraylis.de**