FINBRIDGE

based on competence and commitment

Data Warehousing & Data Vault 2.0

Die zunehmende Digitalisierung in allen Lebensbereichen ist ein Treiber für sich immer schneller verändernde Geschäftsanforderungen, nicht zuletzt in der Finanzwelt. Diese erfordern unter anderem auch eine schnelle Reaktionsfähigkeit von Banken. Im Projektmanagement wird durch agile Methoden darauf reagiert, doch agile Methoden allein sind kein Allheilmittel. Im Data Warehousing ist es wichtig, ein Datenmodell zu finden, welches flexibel genug ist, um den heutigen Anforderungen - regulatorischer wie geschäftsstrategischer Natur - Stand zu halten. So steigen beispielsweise die regulatorischen Anforderungen weiter stetig an und verlangen u.a. eine bessere Nachvollziehbarkeit der ermittelten Kennzahlen. Somit muss auch die Auditfähigkeit innerhalb des DWHs sichergestellt werden.

In diesem Artikel beleuchten wir das klassische Data Warehousing mit einem dimensionalen Modell und Data Vault 2.0 und stellen beide Ansätze gegenüber.

Klassisches Data Warehouse mit einem dimensionalen Modell

Ein Data Warehouse dient als zentrale Datenschnittstelle der IT Architektur, die - zumindest theoretisch - alle Datenströme der Quellsysteme aufnimmt, integriert, transformiert, historisiert und bereitstellt. Der klassische Aufbau eines DWHs ist eine 3-Schicht-Architektur, welche aus folgenden drei Layern besteht:

- Staging Area (Eingangsschicht)
- Core Data Warehouse (Integrationsschicht)
- Data Mart (Bereitstellungsschicht)

Sowohl die Datenaufnahme als auch die Datenweitergabe zwischen den Layern und an die Abnehmer erfolgt über ETL (Extract, Transform, Load)-Prozesse, vgl. auch Abbildung 1.

Staging Area Layer

Nach der Extraktion aus den Quellsystemen werden die Daten unverändert in der Staging Area abgelegt. Die Staging Area dient als temporärer Zwischenspeicher, um die operativen Systeme zu entlasten. Dies bietet außerdem den Vorteil, dass auf der Staging Area SQL-Statements angewendet werden können, was z.B. bei Eingangsdaten in Form von Flat Files nicht möglich ist.

Core Data Warehouse Layer

Im ETL-Prozess von der Staging Area zum Core DWH findet die Integration, Konsolidierung und Ableitung von Businesslogik statt. Dies geschieht, indem die Daten transformiert und in Dimensions- und Faktentabellen geschrieben werden.

Ein klassisches Datenmodell eines dimensionalen DWHs stellt das Star Schema dar. Neben dem Star Schema gibt es auch die sogenannten Galaxy- und Snowflake Schemen, die wiederum eine Erweiterung des zuerst genannten Datenmodells darstellen.

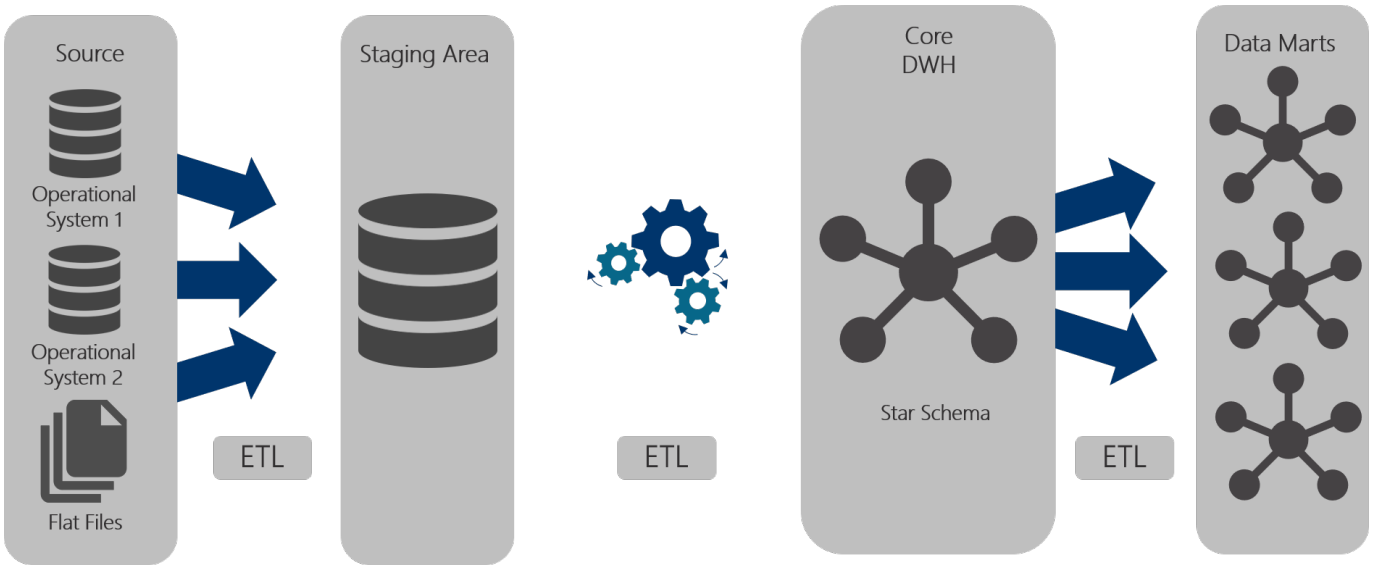

Abbildung 1: *Architektur im dimensionalen Modell.*

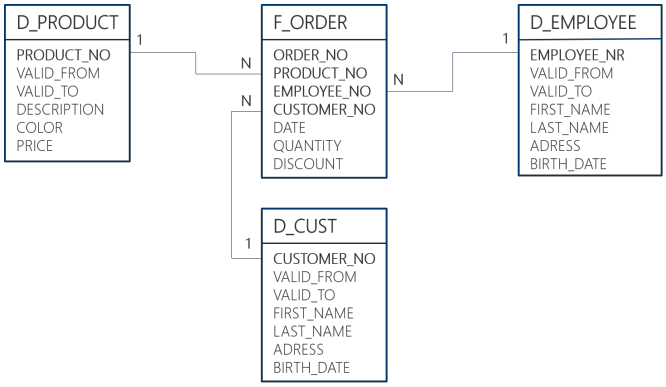

Abbildung 2: *Beispiel dimensionales Modell/Star Schema: Die Dimensionstabellen (hier z.B.:* D_PRODUCT*) sind sternförmig um die Faktentabellen (hier:* F_ORDER*) angeordnet.*

In diesem Artikel werden wir das Star Schema näher beleuchten, siehe auch Abb. 2. Im Star Schema sind die Dimensionstabellen sternförmig um eine Faktentabelle angeordnet und beinhalten die beschreibenden Attribute. Die Faktentabellen enthalten neben den Fremdschlüsseln (z.B. die Kundennummer) die für die Geschäftsprozesse relevanten Messgrößen und Kennzahlen, wie beispielsweise Kreditsummen oder Mengen. Neben dem Inhalt unterscheiden sich Dimensions- und Faktentabellen durch die Beladeprozesse. Häufig werden die Dimensionen nach der Methode “Slowly Changing Dimensions” Typ 2 (SCD Typ 2) beladen, siehe auch Abbildung 3.

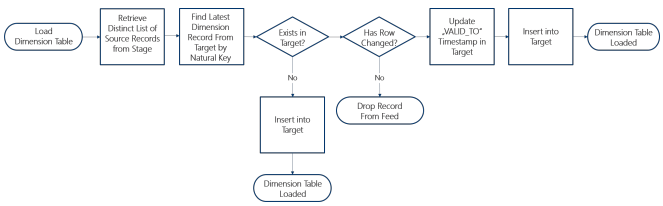

Abbildung 3: *Beladung der Dimensionstabellen nach SCD Typ 2.*

Ein wesentlicher Bestandteil dieser Methode ist, dass alle Datensätze Gültigkeitszeiträume erhalten (gültig von/ gültig bis). Dabei wird im ersten Schritt geprüft, ob der Schlüssel des eingehenden Datensatzes bereits in der Tabelle vorhanden ist. Ist dies der Fall, so wird ein Update auf dem bestehenden Datensatz durchgeführt und dieser erhält einen neuen Gültigkeitszeitraum (z. B. gültig bis gestern). Danach wird der neue Datensatz in die Tabelle geschrieben.

Bei der Beladung von Faktentabellen wird zunächst die referentielle Integrität geprüft. Dabei muss jeder Fremdschlüssel in den korrespondierenden Dimensionen vorhanden sein. Ist dies der Fall, so wird der neue Datensatz in die Faktentabelle geschrieben.

Data Mart Layer

Nach der Beladung des Core DWH erfolgt die Bereitstellung im Data Mart Layer. Im Data Mart Layer werden Teildatenbestände auf die jeweilige Nutzung abnehmerspezifisch zugeschnitten. Je nach Anwendungsfall können die Abnehmer auf einen Data Mart oder direkt auf das Core DWH zugreifen.

Data Warehouse mit Data Vault 2.0

Data Vault beschreibt die Methodik, Architektur und Modellierung moderner DWHs. Methodisch wird dabei auf das Process Decision Framework “Disciplined Agile Delivery”zurückgegriffen. Im Folgenden stellen wir Ihnen die Architektur und das Modell vor.

Ähnlich der eingangs beschriebenen klassischen Architektur, gibt es auch bei Data Vault einen 3-Schicht-Aufbau wie in Abbildung 4 illustriert. Während die Staging Area analog zur Staging Area des klassischen DWHs aufgebaut ist, wird im Core DWH zusätzlich zwischen Raw Vault und Business Vault unterschieden. Auch in der Bereitstellungsschicht wird zwischen Raw Mart und Information Mart unterschieden.

Staging Area Layer

Die Staging Area ist quellsystemorientiert aufgebaut. Die Funktion der Staging Area ist identisch zum klassischen DWH mit einem dimensionalen Modell. Bei der Beladung der Staging Area werden sogenannte “Hard Business Rules”angewendet. Eine Faustregel für Hard Business Rules ist: Der Inhalt und die Bedeutung der Daten wird nicht verändert, lediglich die Art, wie die Daten gespeichert werden, ändert sich. Ein typischer Anwendungsfall für Hard Business Rules wäre z. B. die Anlieferung eines Flat Files in Form einer CSV-Datei. Die Daten in einer CSV-Datei sind untypisiert. Beim Beladen der Staging Area werden die Daten 1:1 übernommen. Zusätzlich werden die Daten um Datentypen ergänzt.

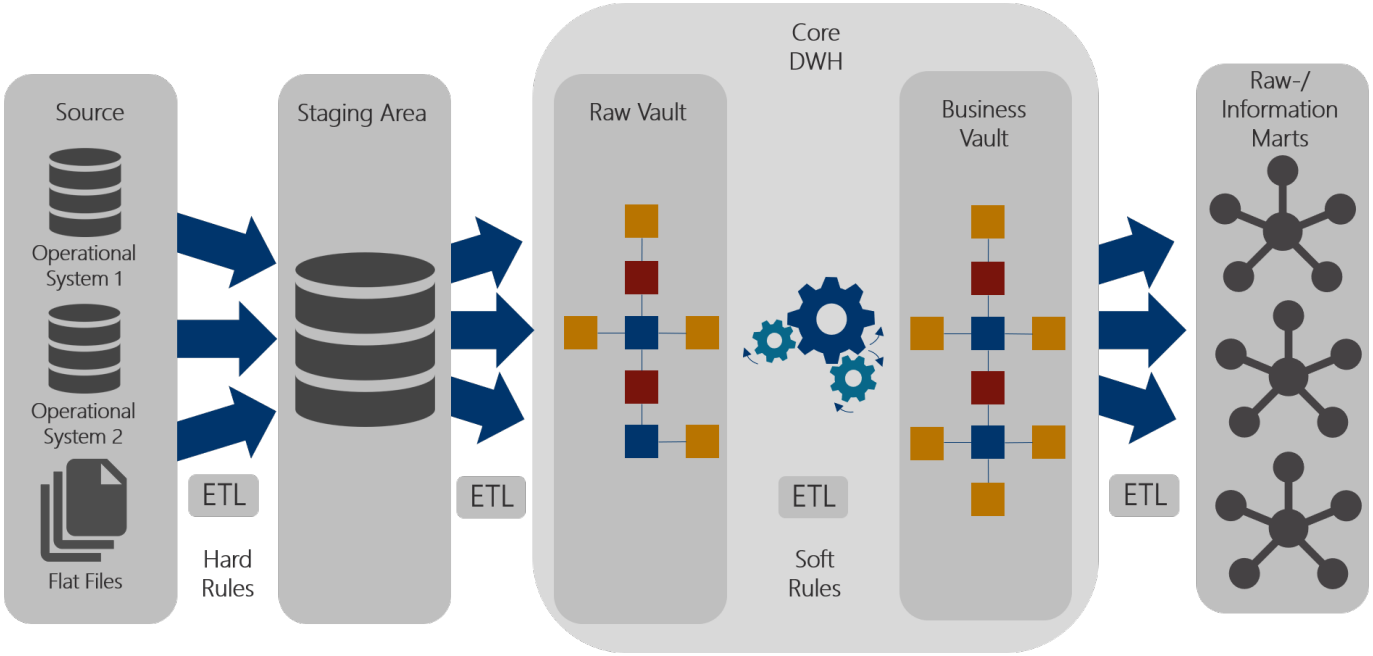

Abbildung 4: *Architektur Data Vault 2.0.*

Core Data Warehouse Layer

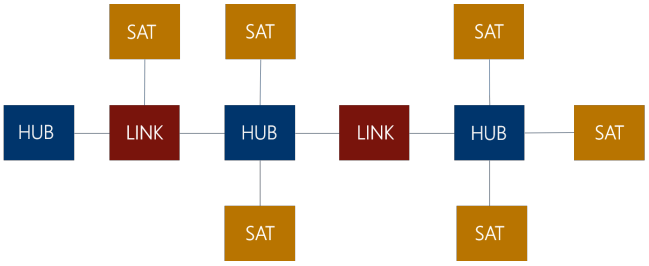

Abbildung 5: *Data Vault Modell für den Core DWH Layer: Die Datentabellen werden unterschieden in Hubs (HUB), Satelliten (SAT) und Links (LINK).*

Beim Übergang von der Staging Area zum Raw Vault bleiben die Daten weiterhin unverändert. Im Raw Vault findet die Historisierung und Integration statt. Dies geschieht, indem die Daten auf Hubs, Satelliten und Links aufgeteilt werden, vgl. Abbildung 5. Dabei beinhalten die Hubs Schlüssel (Business Keys), die eine Entität eindeutig identifizieren. Dies können z.B. Kontonummern oder Kundennummern sein. In den Satelliten werden beschreibende Informationen gespeichert. Links bilden die Beziehungen zwischen den Entitäten ab und verbinden somit mehrere Hubs miteinander. Links können aber auch Hubs mit Satelliten verbinden, wenn die Satelliten eine Beziehung beschreiben.

Hubs, Satelliten und Links haben folgende Standardfelder: Load Date Timestamp, Recordsource und einen Hash Key. Satelliten können zusätzlich noch einen Hash Diff (Hash Difference) haben. Der Load Date Timestamp beschreibt den Zeitpunkt, zu dem das Element in die Tabelle geladen wurde. Die Recordsource beschreibt, aus welchem Quellobjekt die Daten stammen. Hash Keys sind künstliche Schlüssel und dienen zur Identifizierung eines Hub- oder Link-Elementes. Der Hash Diff beinhaltet alle beschreibenden Felder eines Satelliten und ermöglicht so eine schnelle Erkennung von Differenzen.

Alle Tabellen werden durch Insert-Only-Prozesse beladen. Abbildung 6 stellt beispielhaft den Beladungsprozess eines Satelliten dar. Bei Links und Hubs wird lediglich geprüft, ob der Hash Key bereits vorhanden ist. Ist dies nicht der Fall, so wird der neue Datensatz geschrieben. Bei Satelliten wird zusätzlich der Load Date Timestamp und der Hash Diff geprüft.

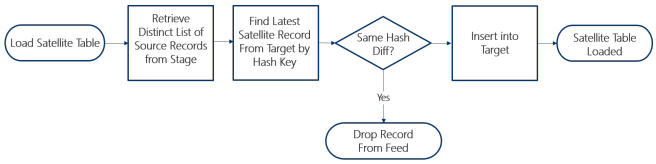

Abbildung 6: *Beladung von Satelliten.*

Nach der Beladung des Raw Vaults, erfolgt die Beladung des Business Vaults. Der Business Vault befindet sich im selben Schema wie der Raw Vault. Aufgabe des Business Vaults ist die Konsolidierung und Abbildung von Business Rules. Bei der Beladung des Business Vaults kommen sogenannte “Soft Business Rules”zum Einsatz. Soft Business Rules bilden die Geschäftsanforderungen ab und verändern den

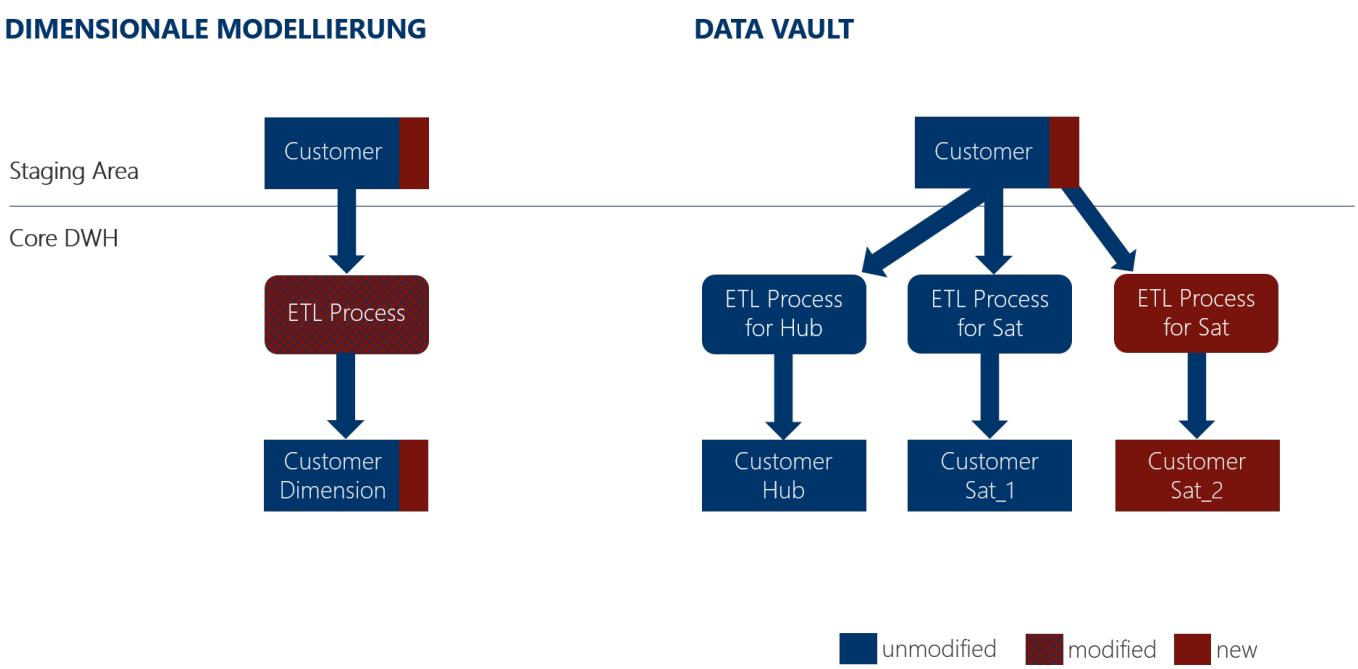

Abbildung 7: *Beispiel Erweiterung des Modells [Rei16].*

Inhalt und die Bedeutung der Daten. Beispielsweise werden aus angelieferten Kennzahlen KPIs (Key Performance Indicators) abgeleitet, Aggregationen in Form von Gruppierungen wie z. B. Altersgruppen durchgeführt oder die Konsolidierung von Daten aus unterschiedlichen Systemen vorgenommen.

Besonders hervorzuheben ist, dass alle Anpassungen am Data Vault Datenmodell nur additiv durchgeführt werden, siehe auch Abbildung 7. Werden beispielsweise neue Attribute an das Core DWH bereitgestellt, so werden neue Satelliten erstellt. Bestehende Tabellen und ETL-Prozesse bleiben unberührt.

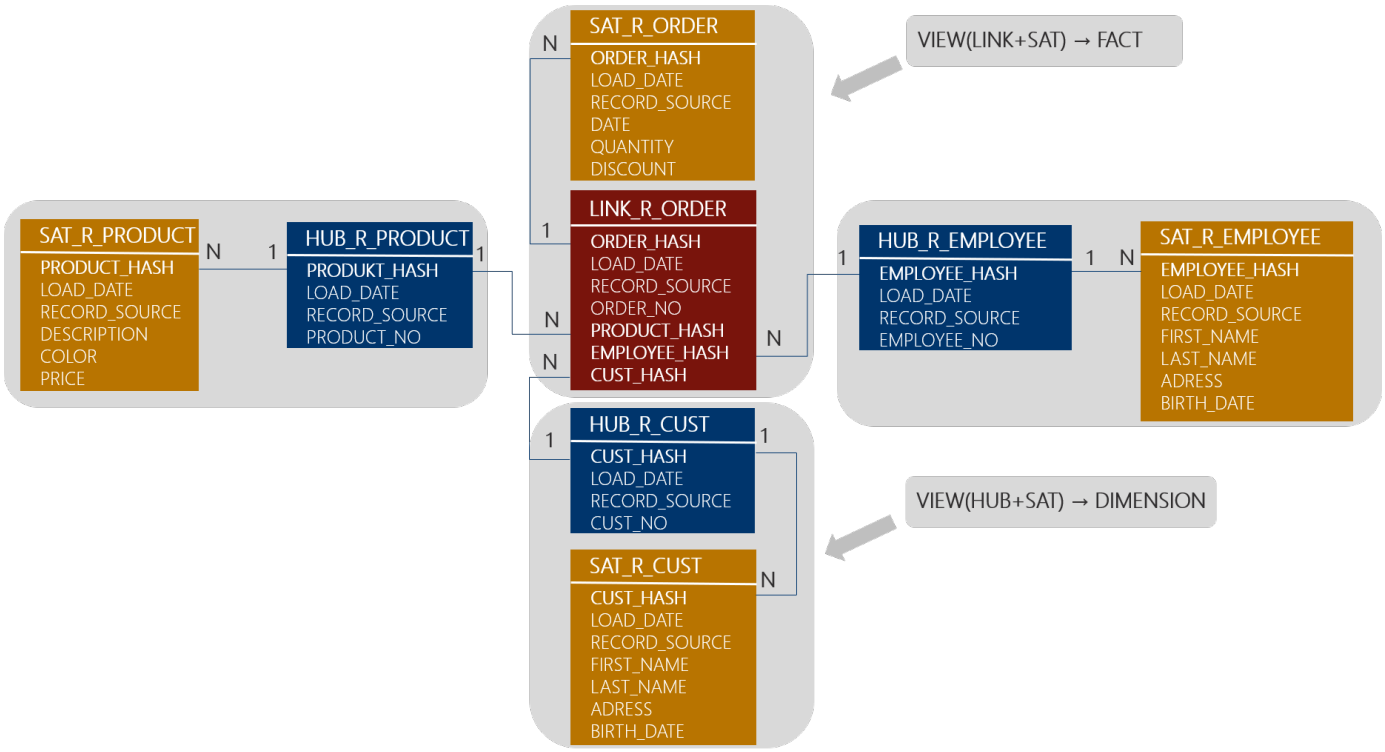

Abbildung 8: *Beispiel Views für einen Raw Mart [Mue17].*

Raw-/Information Mart Layer

Der Zugriff auf die Daten erfolgt ausschließlich über den Raw- bzw. Information Mart Layer, beispielhaft in Abbildung 8 dargestellt. Es wird dabei zwischen Raw Mart und Information Mart unterschieden. Ein Raw Mart beinhaltet nur Daten aus dem Raw Vault. Ein Information Mart beinhaltet sowohl Daten aus dem Raw Vault als auch Daten aus dem Business Vault. Im Raw- bzw. Information Mart Layer können die Daten abnehmerspezifisch bereitgestellt werden. Typisch ist die Bereitstellung eines dimensionalen Modells oder in Form von flachen Tabellen (denormalisierte Tabellen). Das gewünschte Bereitstellungsmodell kann erstellt werden, indem Views über die Tabellen aus dem Data Vault gelegt werden.

Gegenüberstellung

In der folgenden Abbildung 9 stellen wir die Vor- und Nachteile zwischen einem Core DWH mit einer dimensionalen Modellierung und einem Core DWH mit einem Data Vault Modell dar. Bei der Entscheidung, welches Modell man für sein Data Warehouse auswählen sollte, gibt es, wie so oft, kein richtig oder falsch. Die richtige Kombination aus dimensionaler Modellierung und Data Vault kann der Schlüssel zum Erfolg sein.

Das dimensionale Modell ist besonders gut geeignet, wenn ein klar definiertes Zielbild vorhanden ist und das Ergebnis am Ende möglichst statisch bleibt. Daher bietet sich dieser Modellierungsansatz vor allem für Data Marts an. Hier punktet das dimensionale Modell mit seiner einfachen Nachvollziehbarkeit.

Ist das Zielbild hingegen nicht ganz klar bzw. befindet sich noch in der Konkretisierung, da die an das DWH anzubindenden Systeme noch in Klärung sind oder die Umsetzung regulatorischer Anforderungen noch nicht final definiert ist, so bietet sich das Data Vault Modell auf Grund seiner Flexibilität an.

Wie unterstützt Finbridge seine Kunden?

Mit unserer fachlichen Expertise und technischem Knowhow unterstützen wir Sie gerne bei der Einführung und Weiterentwicklung Ihres DWHs in folgenden Bereichen:

- Projektmanagement
- Anforderungsmanagement
- Fach- und DV-Konzeption
- Implementierung
- Testmanagement und -durchführung
- Strategische Beratung bei der Auswahl eines geeigneten Modells.

	DIMENSIONALE MODELLIERUNG	DATA VAULT MODELL
VORTEILE	– Für Business User einfach nachzuvollziehen. – Ideale Struktur für Reporting-Anwendungen. – Schnelle Leseprozesse durch wenige Joins.	– Einfache Integration neuer Daten, da bestehende Tabellen und ETL-Prozesse unberührt bleiben. – Hohes Automatisierungspotential durch standardisierte Prozesse und Modellierungskonventionen. – Geringer Testaufwand, da nur die Erweiterungen getestet werden müssen. – Hohes Parallelisierungspotential aufgrund geringer Abhängigkeiten innerhalb der Beladung. – Vollumfängliche Historisierung
NACHTEILE	– Modellerweiterung aufwändig, da bestehende Tabellen und ETL-Prozesse anzupassen sind. – Erhöhter Testaufwand (neue & alte Funktionalitäten) – Komplexere Anhängigkeiten, da häufig mehrere Systeme in die gleichen Dimensions- und Faktentabellen schreiben. Dies erfordert die Überprüfung der referentiellen Integrität. – Schreibprozesse erfordern aufwändigere Update- und Delete-Operationen. – Wenig Automatisierungspotenzial – I.d.R. keine vollumfängliche Historisierung wg. frühzeitiger Filterungen und Aggregationen vor dem DWH	– Langsame Leseprozesse aufgrund vieler Joins. – Struktur im Core DWH ist nicht geeignet für den direkten Zugriff durch Reporting-Anwendungen. – Geringe Abweichungen vom Modellierungsstandard können sich schnell zu großen Nachteilen entwickeln.

Abbildung 9: *Dimensionale Modellierung vs. Data Vault Modell. Quelle: Finbridge GmbH & Co. KG*

Kontakt

FINBRIDGE
based on competence and commitment

Frank Kirr
Expert Consultant

Finbridge GmbH Co. KG
Louisenstraße 100
61348 Bad Homburg
www.finbridge.de

Telefon: +49 6172 499770
Telefax: +49 6172 49977-11
Mobil: +49 151 58258996
frank.kirr@finbridge.de

FINBRIDGE
based on competence and commitment

Ilja Jost
Senior Consultant

Finbridge GmbH Co. KG
Louisenstraße 100
61348 Bad Homburg
www.finbridge.de

Telefon: +49 6172 499770
Telefax: +49 6172 49977-11
Mobil: +49 151 58259015
ilja.jost@finbridge.de

Quellen

LiO15 Dan Linstedt/ Michael Olschimke “Building a Scalable Data Warehouse with Data Vault 2.0”, 2015.

Mue17 Michael Müller: “Data Vault: Modellierungsansatz für ein Data Warehouse”, 20. April 2017, Link zum Artikel.

Rei16 Reinhard Mense: “BI-Agilität durch Zusammenspiel”. In: BI-Spektrum 03/2016, S. 36-39.