Data Vault – Modellierungsmethode für agile Data Warehouse Systeme

Dr. Bodo Hüsemann
Informationsfabrik GmbH
Münster

Schlüsselworte

Data Vault, Datenmodellierung, Data Warehouse, Agile Vorgehensweisen

Einleitung

Die Kern-Technologien und Standard-Vorgehensweisen zur Implementierung von Data Warehouse Systemen sind langjährig erprobt und in der Praxis bewährt. Auf der anderen Seite sind aktuelle Anforderungen hinsichtlich kürzerer Aktualisierungsintervalle, höherer Ladeperformance bei wachsender Datenmenge und gleichzeitig agiler Adaptierbarkeit mit den traditionellen Methoden oft nicht mehr effizient umsetzbar. Hier müssen neben neuen technologischen Pfaden (z.B. InMemory-Datenbanken, Big Data Technologie) auch konzeptionell und methodisch neue Wege beschritten werden.

In den letzten Jahren hat sich mit "Data Vault" eine neue Data Warehouse Modellierungsmethode profiliert, die speziell für die Erfordernisse eines Data Warehouse entwickelt wurde und für Integrationsschichten eine Alternative zu klassischen Modellierungsmethoden (ER-Modellierung, 3NF) darstellt. Data Vault wurde in seinen Grundzügen von Dan Linstedt bereits vor über 10 Jahren entwickelt und hat seine Feuerprobe in den USA bei zahlreichen Großprojekten erfolgreich bestanden. Auch im deutschsprachigen Raum stellen prominente Projekte die Leistungsfähigkeit eines Data Vault basierten Data Warehouse Systems unter Beweis.

Der Lösungsansatz bietet

- ein einfaches Basismodell mit wenigen Grundkonzepten
- Struktur-Entkopplung und Impact-Isolation für Modelländerungen und -erweiterungen
- massiv parallelisierbare Ladeprozesse mit Realtime-Unterstützung
- flexible Strukturerweiterung bei gleichzeitiger Historisierungsoption

Grund genug also, um in diesem Beitrag die wesentlichen Grundbausteine vorzustellen, die grundsätzliche Vorgehensweise für ETL Prozesse zu skizzieren und die Hauptargumente für den Einsatz zu erläutern.

Data Vault Modellierungsmethode

Aus relationaler Sicht werden in herkömmlichen Datenmodellen innerhalb einer Tabelle Schlüssel-, Relations- und Kontextinformation gemeinsam gespeichert und historisiert. Das Data Vault Datenmodell entkoppelt diese Bestandteile in gesonderte Teil-Entitäten (vgl. im Folgenden Abb. 1).

Ein Data Vault **Hub** (Notation: blau) enthält die Zuordnung eines fachlichen Schüssels zu einem künstlichen Schlüssel (engl. surrogate key und gleichzeitig primary key, PK). Ein fachlicher Schlüssel

ist das identifizierende Merkmal (im Beispiel „CustomerNo“) einer konzeptionellen Entität (z.B. Kunde, Vertrag, Produkt). Als Integrationsmerkmal für ein Data Warehouse sollte ein fachlicher Schlüssel die folgenden Eigenschaften erfüllen: Eindeutigkeit, Unveränderlichkeit, Applikationsneutralität und prozessübergreifende Kommunikationsfähigkeit (z.B. in Berichten). Die Attribute in einem Hub sind unveränderlich (Schlüsselinformation ist invariant). Die fachliche Gültigkeit von Kontext (historisierte Attribute, z.B. „Name“) wird in zugeordneten Satelliten festgehalten (s.u.).

Die richtige Definition von fachlichen Schlüsseln ist der Anker eines guten Data Vault Modells. Daher sollte der Analyse in diesem Bereich große Aufmerksamkeit geschenkt werden, damit ein langfristig tragfähiges Fundament zur Datenintegration entsteht.

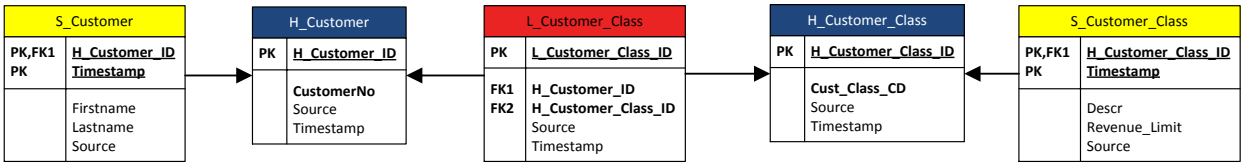

Abb. 1: Logisches Data Vault Datenmodell Kunde.

Ein Data Vault **Link** (Notation: rot) repräsentiert eine natürliche Beziehung/Relation zwischen zwei oder mehreren fachlichen Hauptentitäten und besitzt selbst keine eigenständige Identität außer dem zusammengesetzten Schlüssel der verbundenen Entitäten. Ein Link assoziiert somit eine Menge Hubs oder Links und repräsentiert Relationen unabhängig von ihrer Kardinalität (1:1, 1:n, n:m). Damit ist sichergestellt, dass ein Link vergangene, derzeitige und zukünftige Beziehungen abbilden kann, ohne dass eine aufwendige Reorganisation notwendig ist (z.B. bei fachlicher Änderung der Kardinalität). Neben der Relation werden innerhalb eines Links keine weiteren Informationen hinterlegt und Attribute eines Links sind ebenfalls unveränderlich. Damit ergibt sich eine konzeptionelle Analogie zwischen Links und Hubs, die beide zur fachlichen Identifikation einer Entität/Relation und der Zuordnung eines künstlichen Schlüssels dienen.

Die beschreibenden Attribute eines Hubs/Links werden in abhängigen Data Vault **Satelliten** (Notation: gelb) gespeichert. Satelliten speichern fachliche Information (z.B. den Kundennamen) getrennt von der zugeordneten Schlüsselinformation des Hubs/Links. Der Schlüssel eines Satelliten ist zusammengesetzt aus dem künstlichen Schlüssel des zugeordneten Hubs/Links und einem Zeitkontext (z.B. ein Gültigkeitsintervall VALID_FROM, VALID_TO). Das heißt die Änderung der Attributwerte wird in Satelliten historisiert gespeichert und gewährleistet damit die volle Auditierbarkeit eines Data Vault basierten Data Warehouse. Pro Hub/Link können mehrere Satelliten existieren, die Attribute nach unterschiedlichen Kriterien zusammenfassen, z.B. weil sie

- einen fachlichen Zusammenhang darstellen (Gruppierung nach dem Fachkontext)
- bei Änderungen gemeinsam betroffen sind (Gruppierung nach Historisierungskontext)
- aus der gleichen Quelle geladen wurden (Gruppierung nach Datenherkunft)
- die gleichen Datentypen besitzen.

In herkömmlichen 3NF Datenmodellen werden in Data Warehouse Projekten bestehende Strukturen mit Fremdschlüsseln für neue 1:n Relationen erweitert oder neue Attribute in existierenden Tabellen hinzugefügt. Ein Data Vault Datenmodell ermöglicht durch die Aufspaltung des Datenmodells in Hubs, Links und Satelliten die schrittweise Erweiterung ohne bestehende Teile zu verändern. So

können beispielsweise durch neue Links in Verbindung mit Hubs zusätzliche fachliche Entitäten hinzugefügt werden (z.B. könnte in Abb. 2 der Satellit „S_Emp_Salary“ oder die Einzelpositionen eines Verkaufs in einem zweiten Inkrement entstanden sein). Außerdem können z.B. bei der Anbindung von neuen Datenquellen durch neue Satelliten für bestehende Hubs weitere Attribute hinzugefügt werden. Ein Data Vault Datenmodell unterstützt also agile, inkrementelle Entwicklungsmethoden und beschränkt durch Kapselung von Änderungen in neuen Entitäten den Aufwand für Impactanalysen auf das notwendige Minimum.

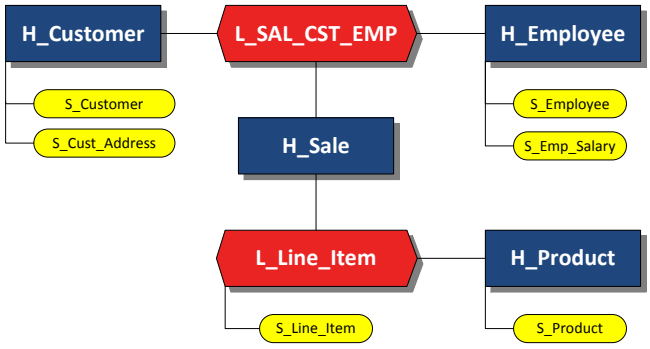

Abb. 2: Konzeptionelles Data Vault Datenmodell Kunde – Mitarbeiter - Verkauf.

ETL-Prozesse

Das Data Vault Datenmodell ermöglicht durch die Trennung der Strukturen eine massiv parallele Beladung des Data Warehouse. Im Gegensatz zu den komplexen Ladenetzen herkömmlicher Data Warehouse Systeme existieren im Data Vault Datenmodell keine transitiven Abhängigkeiten, die zu langen sequentiellen Ladeketten führen. Grundsätzlich werden in einem Data Vault Warehouse innerhalb eines dreistufigen Prozesses alle Hubs, Links und schließlich Satelliten parallel beladen.

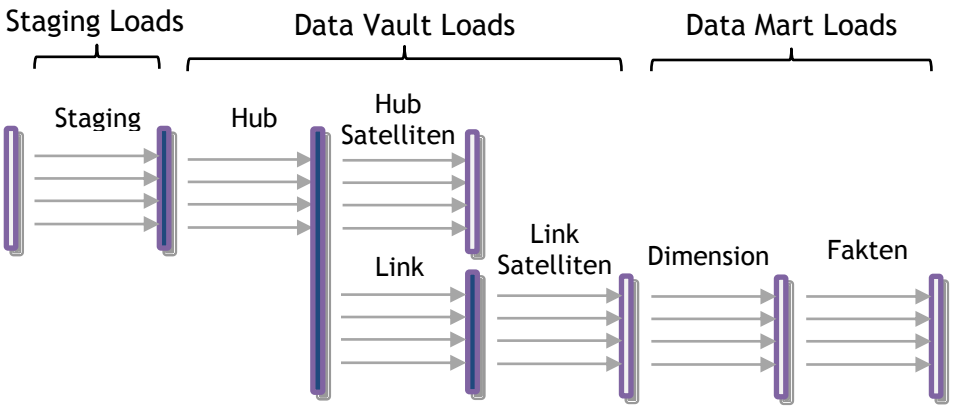

Abb. 3: Data Vault Warehouse ETL Ladeprozess.

Die ETL Prozesse eines Data Vault DWH realisieren für das zentrale Enterprise DWH keine Business Logik (s. folgendes Kapitel zur Architektur) und implementieren eine standardisierte Integrations- und Historisierungslogik. Dies eröffnet den Weg zu stark vereinfachten Laderoutinen mit hohem Generalisierungspotential, so dass auf der Basis einer Metadatenbank über einen Template-Ansatz

oder Generierungs-Verfahren ein hoher Automatisierungsgrad und schnelle Implementierungszeiten bei der Erstellung von ETL Prozessen erreicht wird.

Architektur

In einem traditionellen Data Warehouse System werden in der zentralen Datenhaltungsschicht (im sog. Enterprise Data Warehouse, EDWH) die entscheidungsrelevanten Daten als unternehmensweite Geschäftssicht auf die Daten der operativen Systeme verstanden. Oft wird hier von einem sog. „Single Point of Truth“ gesprochen, also dem erklärten Ziel der Vereinheitlichung und Zentralisierung von Geschäftslogik bei der Befüllung des EDWH. Es soll hierbei abnehmenden Berichtssystemen oder Data Marts ein „normalisierter“, bereinigter Datenbestand zur Verfügung stehen und somit die Konsistenz von Auswertungen in unterschiedlichen Analysebereichen gewährleistet werden (vgl. Abb. 4).

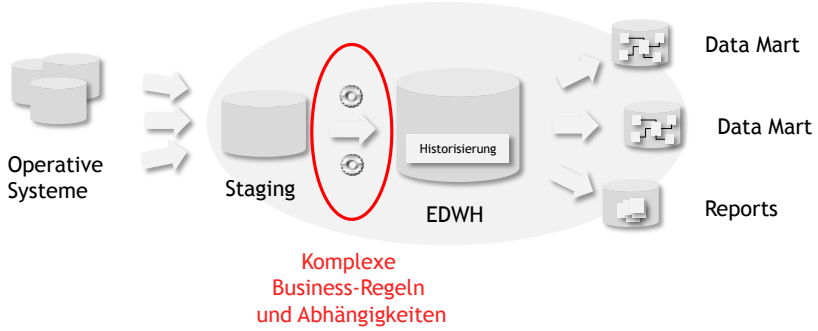

Abb. 4: Standard Data Warehouse Architektur Skizze.

In einem Data Vault Data Warehouse System wird diese Vorgehensweise weiter verfeinert, indem zunächst die Datenintegrationslogik und Historisierungsfunktion von der Geschäftslogik getrennt werden. Im sogenannten „Raw Data Vault“ werden alle auswertungsrelevanten operativen Datenquellen in einer gemeinsamen Historisierungsschicht gespeichert. Das Datenmodell der Raw Vault orientiert sich an den Quellsystemen und hat zum Ziel, die Daten der Quellsysteme möglichst unverfälscht und vollständig zu historisieren und in einem gemeinsamen Datenmodell zu integrieren. Insofern wird im Kontext einer Raw Vault häufig von einer „Single Version of Facts“ gesprochen, also einer revisionssicheren Faktensammlung, die exakt die Historie der operativen Systeme widerspiegelt.

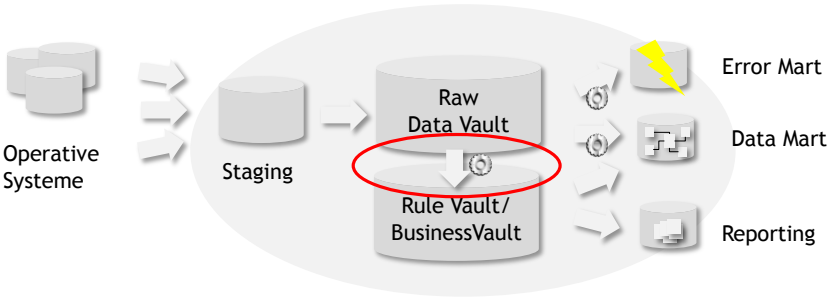

Abb. 5: Data Vault Warehouse Architektur Skizze.

In einer „Rule Vault“ findet aufbauend auf den integrierten Daten der Raw Vault die zentralisierte Anwendung von Geschäftslogik (die fachliche Interpretation oder Bereinigung) und die Transformation in ein fachliches Datenmodell (sog. „Business Vault“) statt. Auf dieser Basis können nun abnehmende Systeme, je nach fachlicher Anforderung, auf die zentralisierte Geschäftslogik der

Rule Vault zugreifen oder, falls erforderlich, auf die Basisdaten der operativen Systeme innerhalb der Raw Vault, um eine Data Mart spezifische Geschäftslogik zu verwenden. Mit dieser Vorgehensweise besteht die Flexibilität, auf spezielle Anforderungen bestimmter Fachbereiche beim Aufbau von Data Marts einzugehen und allgemein die zentrale Unternehmenssicht für alle Abnehmer bereitzustellen.

Insgesamt wird in einem Data Vault System also die Geschäftslogik hinter die Integrations- und Historisierungslogik verlagert. Dadurch können neue Systeme schnell an die Raw Vault angebunden werden, ohne dass notwendiger Weise Geschäftslogik und abhängige Data Marts hiervon betroffen sind. Trotzdem stehen diese Daten aber bereits für neue Auswertungen im Reporting zur Verfügung. Die Adaptionsgeschwindigkeit des Data Warehouse Systems gegenüber neuen Anforderungen aus dem Fachbereich oder Änderungen an den operativen Datenquellen steigt beträchtlich bei gleichzeitiger Kostenreduktion und Risikoverminderung.

Kontaktadresse:

Dr. Bodo Hüsemann

Informationsfabrik GmbH

Scheibenstraße, 117

D-48153 Münster

Telefon: +49 (0) 251-919979 61

Fax: +49 (0) 251-919979 71

E-Mail bhuesemann@informationsfabrik.com

Internet: www.informationsfabrik.com