TDWI Poster

DATA WAREHOUSE AUTOMATION MIT DATA VAULT

tdwi WISSEN

DATA WAREHOUSE AUTOMATION MIT DATA VAULT

DATA VAULT

… bezeichnet eine Modellierungstechnik für das Data Warehouse

… folgt klaren Regeln

… hohe Flexibilität im Hinblick die Erweiterbarkeit

… legt den Fokus darauf, den stetig neuen Geschäftsanforderungen gerecht zu werden

… vereint Aspekte der rationalen Datenbankmodellierung mit der dritten Normalform sowie dem Sternschema

… vollständige Historisierung der Daten

Die Trennung der Konsolidierung und ihrer Business-Logik von der Datenintegration und Historisierung kann ein Teil der Verarbeitung standardisiert und damit automatisiert werden

Quellen: [1, 2]

WEITERENTWICKLUNG ZUM DATA VAULT 2.0

- Die ersten Veröffentlichungen zum Thema Data Vault gehen auf Linstedt um das Jahr 2000 zurück.
- Es folge eine stetige Weiterentwicklung des Datenmodels, in dessen Zusammenhang auch der Begriff **Data Vault 2.0** entstand.
- Diese Erweiterungen wurden von Linstedt in einem Data Vault 2.0 Framework zusammengefasst.
- Es umfasst neben der Modellierung auch die agile Vorgehensweise bei der Implementierung und betrachtet die Architektur.
- Dabei werden zudem Performanceaspekte und unter anderem NoSQL Techniken bis hin zur Automatisierung vertieft.

Quellen: [2, 3, 8]

MODELLIERUNG

Die Grundidee des Data Vault folgt der Aufteilung der Daten in die Strukturelemente HUB, LINK und SATelit.

HUB (Geschäftsobjekt)
- Kernobjekt der Geschäftslogik
- Identifiziert die fachliche Identität
- Speichert:
  - Fachlichen Schlüssel (Business Keys)
  - künstlich generierten Primärschlüssel
  - Quellsystem und Ladezeitpunkt
- Speichert keine beschreibenden Attribute
- Zu beachten: direkte Verbindung zwischen Hubs sind nicht zulässig

LINK (Geschäftsbeziehung)
- Stellt die Beziehungen zwischen den HUBs dar
- Kann beliebig viele HUBs miteinander verknüpfen
- Kann keinen, einen oder mehrere SATteliten verknüpfen

SATteliten
- Informationen, die HUB und LINK beschreiben
- Speichert:
  - Beschreibende Attribute
  - Quellsystem und Ladezeitpunkt
  - Fremdschlüssel zu genau einem HUB oder LINK

Quellen: [2, 4]

MODELLIERUNGSPROZESS

Fachliche Entitäten definieren:
HUBs

↓

Beziehungen modellieren:
LINKs

↓

Beschreibende Attribute festlegen:
SATelites

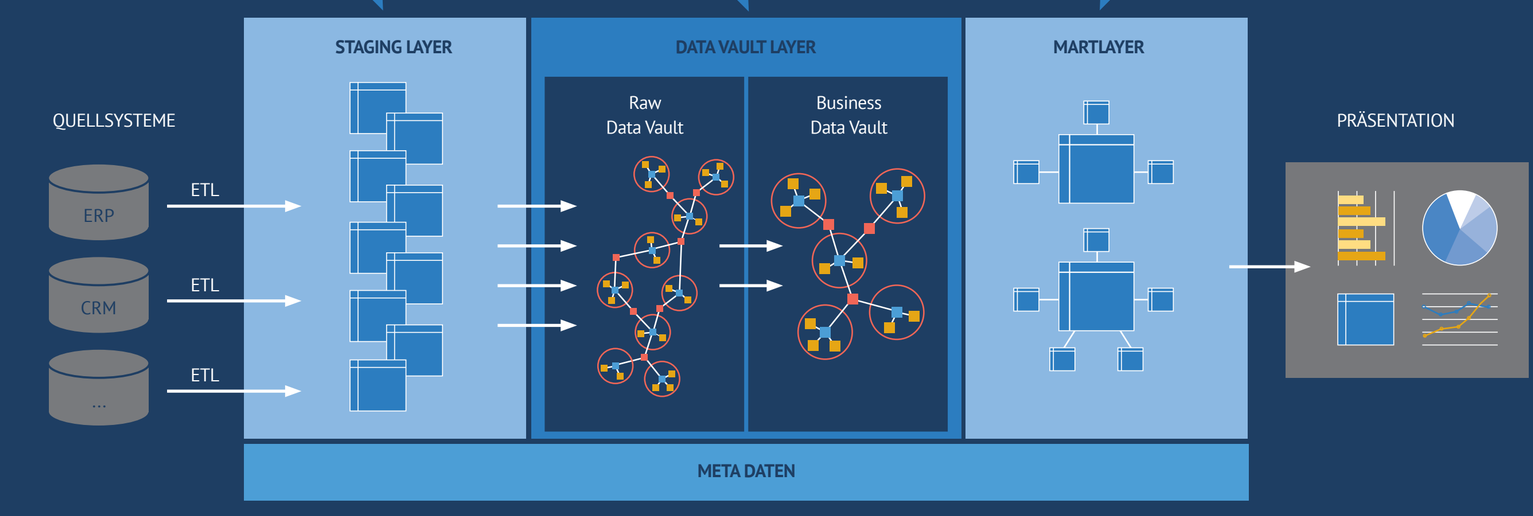
Quellen: [3, 4, 6, 7]

ARCHITEKTUR

Innerhalb des **Staging Layers** werden die Rohdaten aus den verschiedenen Quellsystemen eingesammelt

Hier wird das Data Vault modeliert. Es setzt sich aus den Rohdaten (Raw Data Vault) und harmonisierten und transformierte Daten (Business Data Vault), welche auf Geschäftsregeln und Laufzeitinformationen basieren zusammen

Im Mart Layer werden die Daten in die Form eines Star-Schema und/oder eines anderen Modellierungsverfahren transformiert und gespeichert. Diese Form der Daten stellen die Informationsquelle für die Visualisierung mittels der BI Plattformen dar.



Quellen: angelehnt an [2, 4, 5, 9]

VORTEILE VON DATA VAULT FÜR UNTERNEHMEN

ENTWICKLUNGSZEIT REDUZIEREN
Mithilfe von Data Vault lässt sich die Entwicklungszeit reduzieren. Dabei liegen die Vorteile vor allem bei der Implementierung der Business-Anforderungen.

RETURN ON INVESTMENT
Unternehmen erzielen einen höheren Retrun on Investment (ROI).

SKALIERBARKEIT
Durch den Einsatz von Data Vault steigt die Skalierbarkeit des DWH.

NACHVERFOLGBARKEIT
Unternehmen können alle Daten bis zu ihrem Quellsystem nachverfolgen, wenn der Data Vault Ansatz genutzt wird.

INKREMENTELL AUSBAUBAR
Dank des iterativen und agilen Entwicklungszyklus lässt sich das DWH flexibel ausbauen und erweitern.

HISTORISIERUNG UND TIME TRAVELING
Durch die dauerhafte Speicherung der unveränderten Rohdaten, entsteht der Vorteil des Time Traveling, also stichtagsbezogene Auswertungen aus der Vergangenheit ermöglicht.

Quelle: [3]

GOLDENE REGELN

1. Eine Hub-Tabelle gibt seinen Primary Key grundsätzlich nach außen
2. Beziehungen zwischen Hubs gibt es nur über Links
3. Rekursive Strukturen gibt es nur über Links
4. Eine Link-Tabelle hat mindestens zwei Foreign Keys-Beziehungen
5. Eine Link-Struktur kann eine Surrogatschlüssel-Beziehung aufweisen
6. Eine Link-Tabelle kann eine unbegrenzte Anzahl an Hubs verbinden
7. Eine Beziehung zwischen Links ist erlaubt
8. Eine Link-Tabelle kann eine Satellit-Tabelle haben
9. Ein Satellit kann nur eine Elterntabelle haben
10. Ein Satellit kann nicht irgendeine Fremdschlüsselbeziehungen besitzen, mit Ausnahme des Primärschlüssels zu der Elterntabelle (Hub oder Link).

WANN IST DATA VAULT UNINTERESSANT?

Kleine Menge an Anforderungen
Kurze Zeiträume, Berichte zu bauen und zu liefern
Daten nur einmalig nutzen
Nur ein Quellsystem (oder sehr wenige)
Nur einen Analysten im Unternehmen
Keine notwendigen Audit-Anforderungen zur Datenhistorisierung
Keine notwendige Integration mehrere Data Centers
Keine *near real time* Anforderungen
Keine umfangreichen Batch Data
Keine externe Daten außerhalb der eigenen Kontrolle
Keine Trendanalysen
Kein permanentes Re-Engineering

CHANCEN UND RISIKEN

CHANCEN	RISIKEN
INTEGRATION VON DATEN AUS VERSCHIEDENEN QUELLSYSTEMEN Die Quelldaten werden unter Verwendung gemeinsamer Geschäftsschlüssel, die in Hubs gespeichert sind, integriert. Die erforderlichen Geschäftsattribute werden in separaten Satelliten pro Quellsystem gespeichert. Dies erleichtert die Zusammenführung der Informationen für weitere Berichte.	BUSINESS-KENNTNISSE ERFORDERLICH Um erfolgreich Modelle mit Data Vault erstellen zu können, ist es wichtig, die Geschäftskontexte zu verstehen. Andernfalls ist das Risiko im Data Vault hoch, dass nur die Quelldaten kopiert und historisiert werden.
PARALLELES LADEN VON DATEN AUS VERSCHIEDENEN QUELLSYSTEMEN Es gibt keine vordefinierte Ladereihenfolge, die Daten können unabhängig voneinander in den Datenspeicher geladen werden.	DATA VAULT KENNTNISSE ERFORDERLICH Die Grundprinzipien von Data Vault müssen dem gesamten Projektteam bekannt sein. Die Entwickler müssen die ETL-Muster für die verschiedenen Objekte verstehen.
VOLLSTÄNDIGE HISTORISIERUNG ALLER ATTRIBUTE Die Versionierung aller Attribute in den Satelliten ermöglicht die Rückverfolgbarkeit aller Änderungen in der Vergangenheit und die Extraktion der Daten zu einem bestimmten Zeitpunkt.	GROSSE TABELLENANZAHL Viele Modellerweiterungen können Datenmodelle mit einer hohen Anzahl von Tabellen erstellen (Hubs, Links und Satelliten). Entsprechend steigt auch die Anzahl der ETL-Prozesse.
LEICHTE ERWEITERBARKEIT DES DATENMODELLS Zusätzliche Entitäten oder Attribute, die für neue Anforderungen verwendet werden, werden als zusätzliche Tabellen im Datendepot implementiert. Bestehende Tabellen werden normalerweise nicht geändert. Dies hilft, Datenmigration zu vermeiden.	KOMPLEXE EXTRAKTION AUS DEM DATENDEPOT Während das Laden von Data-Vault-Tabellen sehr einfach ist, kann das Extrahieren der Informationen zum Laden von Data-Marts umfangreicher sein. Für eine gute Leistung können Hilfstabellen erforderlich sein.
EINFACHE UND EINHEITLICHE ETL-MUSTER Das Laden von Hubs, Links und Satelliten erfolgt nach einheitlichen Regeln, die immer gleich aufgebaut sind.	ENTSPRECHENDE GESCHÄFTSSCHLÜSSEL ERFORDERLICH Die Bestimmung der geeigneten Geschäftsschlüssel ist eine der größten Herausforderungen in der Datenverwaltung. Ungeeignete Schlüssel erschweren die Integration verschiedener Quellen und erhöhen die Komplexität beim Laden von Data Marts.

Quelle [5]

Quellen:

[1] Michael Müller, Aktuelle Themen in die unternehmensweite BI integrieren: Architekturen in Data Vault 2.0, 2015, https://www.tdwi.eu/fileadmin/tdwi/ext_wissen/whitepaper/MID_mueller_BIS_DataVault_2015_AktuelleThemen.pdf
[2] it-novum, Mit Data Vault zu mehr Agilität im Data Warehouse, http://sigs.de/TDWI_Wissen/it-novum/whitepaper_data-vault.pdf
[3] Dirk Lerner, Data Vault Modellierung, TDWI Wissen, http://www.sigs.de/tdwi/Infografik/TDWI_Infografik_DataVault_Modelling.pdf
[4] Linstedt, Daniel, and Michael Olschimke. Building a scalable data warehouse with data vault 2.0. Morgan Kaufmann, 2015.
[5] Trivadis AG, Data Vault – The ultimate problem solver?, https://www.tdwi.eu/fileadmin/tdwi/1.0_Wissen/White_Paper/Trivadis_biGENiUS_Data_Vault-The_Ultimate_Problem_Solver.pdf
[6] Schnider, Dani, Modellierung agiler Data Warehouses mit Data Vault, 2015, http://dani.schniderkennel.ch/pdf/DOAG_Data_Vault_Modellierung_Vortrag.pdf https://bigenius.info
[7] Hultgren, Hans, Data Vault Modeling Guide – Introductory Guide to Data Vault Modeling, 2015, https://www.tdwi.eu/fileadmin/tdwi/ext_wissen/whitepaper/centinnium-data-vault-modeling-guide.pdf
[8] Cramer, Oliver; Lerner, Dirk, Neue Wege in der Datenmodellierung - Data Vault heißt die moderne Antwort. 2015. https://www.tdwi.eu/fileadmin/tdwi/ext_wissen/whitepaper/ITGAIN_lerner_BIS_DataVault_2015.pdf
[9] Bauer, Lutz, Mehr Umsetzungs-Geschwindigkeit und Flexibilität für Ihr Data Warehouse Industrialisierter Data Vault, 2015, https://www.tdwi.eu/fileadmin/tdwi/ext_wissen/whitepaper/MT_bauer_BIS_DataVault_2015.pdf.

tdwi.eu

Autor:
Prof. Carsten Felden
Technische Universität Bergakademie Freiberg (Sachsen)