# Metadaten und Data Vault (Meta Vault)

Andreas Buckenhofer
Daimler TSS GmbH
Ulm

Schlüsselworte
Metadaten, Data Vault, Datenmodel, Standards, Namenskonventionen, Generierung, GOODS

## Inhaltsverzeichnis



### Überblick

Die Verwaltung von Code und Infrastruktur ist hoch standardisiert. Versionierungssysteme, Testsuiten oder EntwicklungsGUIs haben sich etabliert und sind in einem Projekt gesetzt. Doch wie sieht es mit Daten aus? Daten werden mittlerweile als wichtigstes Gut auch im produzierenden Sektor angesehen, doch deren Management steht noch in den Anfängen.

Metadatenmanagement ist seit Jahren bekannt, wird nur wenig beachtet und beschränkt sich häufig auf Themen wie Datenmodellierung oder Glossar. Die systematische Erfassung und Verwendung von Metadaten führt auch zu erhöhter Produktivität, wenn z.B. Teile des ETL/ELT-Codes generiert wird.

Unterschieden werden Metadaten im engeren Sinn (technisch bzw. fachlich) sowie im weiteren Sinn (Fehlerdaten bzw. Metrikdaten). Fachliche Metadaten beschreiben die Bedeutung der Daten aus Sicht der Anwender. Dies sind insbesondere Geschäftsdefinitionen, Geschäftsobjekte, Attribute oder Geschäftsregeln. Technische Metadaten umfassen z.B. Die Beschreibung von Quellsystemen, Datenmodelle, Business Keys. Diese Metadaten bilden die Grundlage zur Generierung von Code. Fehlerdaten umfassen Ausgesteuerte Daten sowie Fehlerbeschreibungen. Metrikdaten umfassen Operative Informationen wie Laufzeiten oder Anzahl verarbeiteter Datensätze.

### Namenskonventionen

Die Benennung von Tabellen und Attributen sollte sich an Konventionen halten. Im Folgenden ist eine Empfehlung für DWH Tabellen und Attribute aufgeführt.

Tabellen sollten generell in der Einzahl (singular) benannt werden, z.B. Fahrzeug, Kunde, Produkt.

**Tabellen**

- Alternativ können auch Suffixe verwendet werden, z.B. KUNDE_HUB, KUNDE_SAT. Bei der alphabetischen Anzeige in Tools wie SQL Developer stehen diese Tabellen dann dicht zusammen.

## Tabellen – HUB und LINK

| Tabellentyp | Namenskonvention |
|---|---|
| HUB, Raw Vault | HUB_%NAME% |
| HUB, Business Vault Vault | HUBB_%NAME% |
| | |
| LINK, Standardlink, Raw Vault | LNK_%NAME% |
| LINK, Hierarchisch, Raw Vault | HLNK_%NAME% |
| LINK, Same as, Raw Vault | SLNK_%NAME% |
| LINK, Transactional, Raw Vault | TLNK_%NAME% |
| | |
| LINK, Standardlink, Business Vault | LNKB_%NAME% |
| LINK, Hierarchisch, Business Vault | HLNKB_%NAME% |
| LINK, Same as, Business Vault | SLNKB_%NAME% |
| LINK, Transactional, Business Vault | TLNKB_%NAME% |

## Tabellen - SAT

| Tabellentyp | Namenskonvention |
|---|---|
| SAT, Standardsatellit zu Hub, Raw Vault | HSAT_%**SOURCE**%_%NAME% |
| SAT, Standardsatellit zu Link, Raw Vault | LSAT_%**SOURCE**%_%NAME% |
| SAT, Standardsatellit zu Hub, Business Vault | HBSAT_%NAME% |
| SAT, Standardsatellit zu Link, Business Vault | LBSAT_%NAME% |
| | |
| SAT, Multi-Active Satellit zu Hub, Raw Vault | HMSAT_%**SOURCE**%_%NAME% |
| SAT, Multi-Active Satellit zu Link, Raw Vault | LMSAT_%**SOURCE**%_%NAME% |
| SAT, Multi-Active Satellit zu Hub, Business Vault | HBMSAT_%NAME% |
| SAT, Multi-Active Satellit zu Link, Business Vault | LBMSAT_%NAME% |
| | |
| SAT, Effectivity Satellit zu Hub, Raw Vault | HESAT_%**SOURCE**%_%NAME% |
| SAT, Effectivity Satellit zu Link, Raw Vault | LESAT_%**SOURCE**%_%NAME% |
| SAT, Effectivity Satellit zu Hub, Business Vault | HBESAT_%NAME% |
| SAT, Effectivity Satellit zu Link, Business Vault | LBESAT_%NAME% |

# Tabellen – Rest

| Tabellentyp | Namenskonvention |
|---|---|
| Referenz-Tabellen, Raw Vault | REF_%NAME% |
| Bridge-Tabellen, Business Vault | BRD_%NAME% |
| PIT-Tabellen, Business Vault | PIT_%NAME% |
| | |
| Dimensionen | DIM_%NAME% |
| Fakten, Transactional | FCTT_%NAME% |
| Fakten, Periodic Snapshot | FCTP_%NAME% |
| Fakten, Accumulating Snapshot | FCTA_%NAME% |

**Spalten**

| Tabellenspalte | Namenskonvention |
|---|---|
| **Load Date Timestamp** | LDTS |
| **Load End Date Timestamp (optional)** | LEDTS |
| **Record Source** | RSRC |
| **Last Seen Date** | LSDT |
| **Sub Sequence Identifier** | SSQI |

**Meta Vault**
Der Meta Mart besteht aus mehreren Tabellen, die die Metadaten des DWHs aufnehmen. Alternativ kann auch ein 3rd Party Tool diese Aufgabe übernehmen und die Metadaten bereitstellen.

**Metric Vault**
Der Metric Mart speichert Laufzeitinformationen, die während der DWH-Beladung anfallen.
Aufgaben des Metric Marts:

- Fehleranalyse
- Performanzanalyse

Typische Metriken im Metric Mart sind:

- Timing: Start und Ende des ETL-Laufs
- Performance: Lese- oder Schreib-I/O Durchsatz, Anzahl gelesene oder geschriebene Datensätze pro Sekunde
- Volumen: Gesamtzahl Datensätze Quelle, Gesamtzahl Datensätze Ziel, Gesamtzahl ausgesteuerte Datensätze
- Frequenz: Wie oft werden Daten geliefert bzw. wie oft wird der ETL Lauf ausgeführt?

Diese Daten können je nach eingesetzter Technologie auf verschiedenen Wegen gewonnen werden, z.B. durch Instrumentierung des Codes in PL/SQL oder durch Auslesen der Daten aus dem Repository eines ETL-Tools.

**Dokumentation**
Gesetzliche Bestimmungen schreiben die Erfassung von Speicherung vor. Im Bankenumfeld waren viele Unternehmen nicht in der Lage ein Risikomanagement zu betreiben was zu einer Bankenkrise führte. Der Basler Ausschuss für Bankenaufsicht legte in BCBS239 explizit fest, dass eine Bank integrierte Datentaxonomien und Architekturen errichten muss. Dies umfasst neben Metadaten auch die Nutzung von Identifikatoren und Namenskonventionen.

> 33. A bank should establish integrated data taxonomies and architecture across the banking group, **which includes information on the characteristics of the data (metadata), as well as use of single identifiers and/or unified naming conventions for data** including legal entities, counterparties, customers and accounts
>
> Quelle: BCBS239, http://www.bis.org/publ/bcbs239.pdf; Basler Ausschuss für Bankenaufsicht (Basel Committee on Banking Supervision, BCBS)

BCBS239 erwähnt außerdem

- die Erfassung von Glossaren
- Inventarisierung von Validierungsregeln
- Exception Reports, die Datenfehler ausweisen und erklären

Die Datenschutz-Grundverordnung GDPR schreibt in Verordnung 2016/679 vor, dass natürliche Personen das Recht auf „Vergessenwerden“ haben, d.h. dass personenbezogenen Daten gelöscht und nicht mehr verarbeitet werden dürfen.

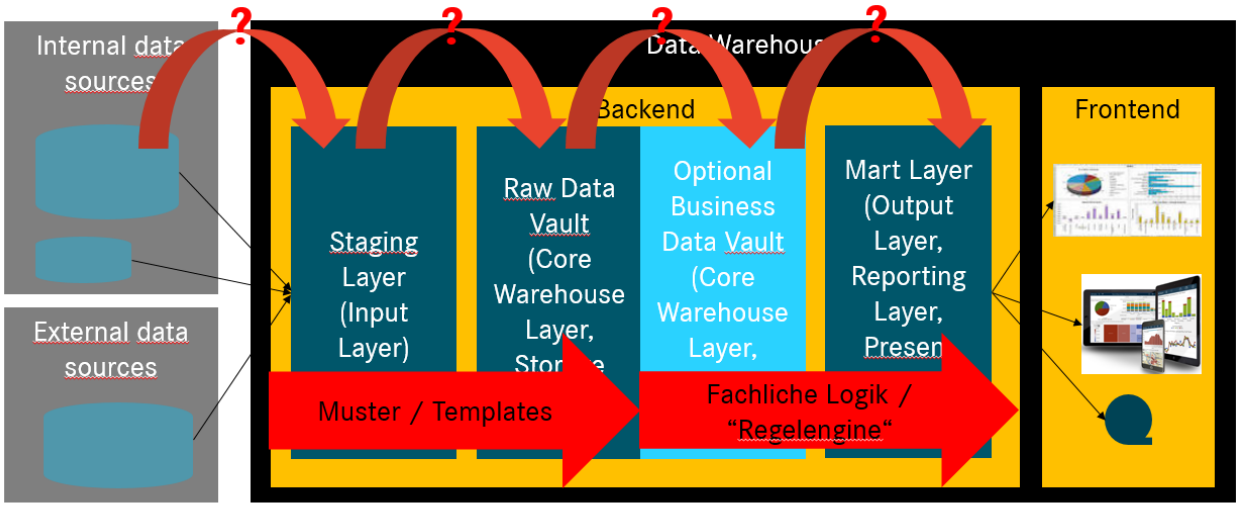
Aufgrund solcher gesetzlichen Bestimmungen steigt die Notwendigkeit einer (toolübergreifenden) Metadatenverwaltung.

**Model-driven Development / Generierung**
Aufgrund der Architektur von Data Vault wird zwischen passiver Datenintegration und Dateninterpretation unterschieden.
Bei der passiven Datenintegration werden verschiedene Quellsysteme mit Hilfe von HUBs / stabilen Business Keys zusammengeführt, d.h. insbesondere

- Keine Datenänderungen
- Keine Datentransformationen
- Keine Datenvereinheitlichungen
- Keine Datenaggregationen

Bei der Dateninterpretation dagegen werden die Daten für den Enduser aufbereitet und fachliche Logik wird angewendet.

Eine Generierung von ETL/ELT-Code eignet sich daher am besten bis (inkl.) zum Raw Data Vault, da die Daten 1:1 transportiert werden. Erst in darauf folgenden Schichte werden Transformationen durchgeführt.






Automatisierung funktioniert nur dann gut, wenn die Ergebnisse aus Automatisierung und manueller Tätigkeit identisch sind. Bei der Generierung von ETL/ELT-Code bis zum Raw Data Vault Layer ist dies gegeben, da Data Vault mit den

- HUB-Tabellen
  Identifikation durch eindeutige natürliche Schlüssel (Business Keys)
- LINK-Tabellen
  Verbindungen zwischen Business Keys (HUBs)
- SAT-Tabellen
  Beschreibende, detaillierte, aktuelle und historisierte Daten

hoch standardisiert ist.

"People have forgotten, or never truly understood, how complex data integration actually is." (Quelle Zitat: http://db2portal.blogspot.de/2015/05/a-trip-report-from-2015-idug-db2-tech.html ). Die Datenintegration ist nach wie vor eine der zeitintensivsten Tätigkeiten beim Aufbau eines DWHs oder auch beim Aufbau einer BigData-Lösung wie Hadoop. Daher ist es von Vorteil, wenn diese Tätigkeiten beschleunigt werden können, insbesondere wenn monotone Kopiertätigkeiten wegfallen, da das Beladen von z.B. HUB-Tabellen immer nach dem gleichen Muster erfolgt.

Data Vault eignet sich für Metadata-driven Development und bietet

- Robuste ETL Prozesse
- Kleine, wartbare Mappings / SQL Statements
- Standardisierbarkeit / Wiederholbarkeit /Automatisierbarkeit
- Flexibilität
- Aber auch:
  Viele Modellierungsoptionen = Viele Diskussionen

Zum Schluss noch ein Blick über den Tellerrand: was macht Google mit Metadaten? Als datengetriebenes Unternehmen hat Google die Wichtigkeit von Metadaten bereits seit Langem erkannt und GOODS (GOOgle Dataset Search) entwickelt. Ein Crawler durchsucht interne Speichersysteme (HDFS/GoogleFS, BigTable, RDBMS) nach Metadaten. Die identifizierten Metadaten werden zentral

gespeichert und ggf. angereichert. Google geht den umgekehrten Weg und hat die Sammlung von Metadaten automatisiert.

Kontaktadresse:
Andreas Buckenhofer
Daimler TSS GmbH
Business Unit Analytics
Wilhelm-Runge-Straße 11
89081 Ulm, Germany
Telefon: +49–(0)731/505–6345
E-Mail Andreas.Buckenhofer@daimler.com
Internet: http://www.daimler-tss.com