Ensemble Modeling am Beispiel von Data Vault

Ben Weissman
@bweissman

Worüber wir heute sprechen

- Was ist ein Ensemble?
- Was ist Data Vault?
- Warum will ich das?
- Warum will ich das nicht?
- Wie kann das überhaupt funktionieren?
  (hier mogelt der Ben dann Biml rein)

Wer bin ich?

GENESEE ACADEMY World-Class Training
Certified Data Vault Modeler

FRIEND OF redgate 2018

Microsoft CERTIFIED
Solutions Expert
Data Management and Analytics

- Ben Weissman, Solisyon, Nürnberg
- @bweissman
- b.weissman@solisyon.de
- SQL Server seit Version 6.5

Please Talk Data To Me

MVP Microsoft® Most Valuable Professional

Data Science — MICROSOFT PROFESSIONAL PROGRAM
Big Data — MICROSOFT PROFESSIONAL PROGRAM

BimlHero CERTIFIED EXPERT

Was ist ein Ensemble? Gibt‘s das nur bei Data Vault?

- Ein Ensemble ist ein Hub mit allen zugehörigen Satelliten und Links
- Hä?
- Gleich!
- Ensembles gibt es nicht nur bei Data Vault, sondern auch bei anderen Methoden wie:
  - Focal Point
  - Anchor
  - Hyper Agility

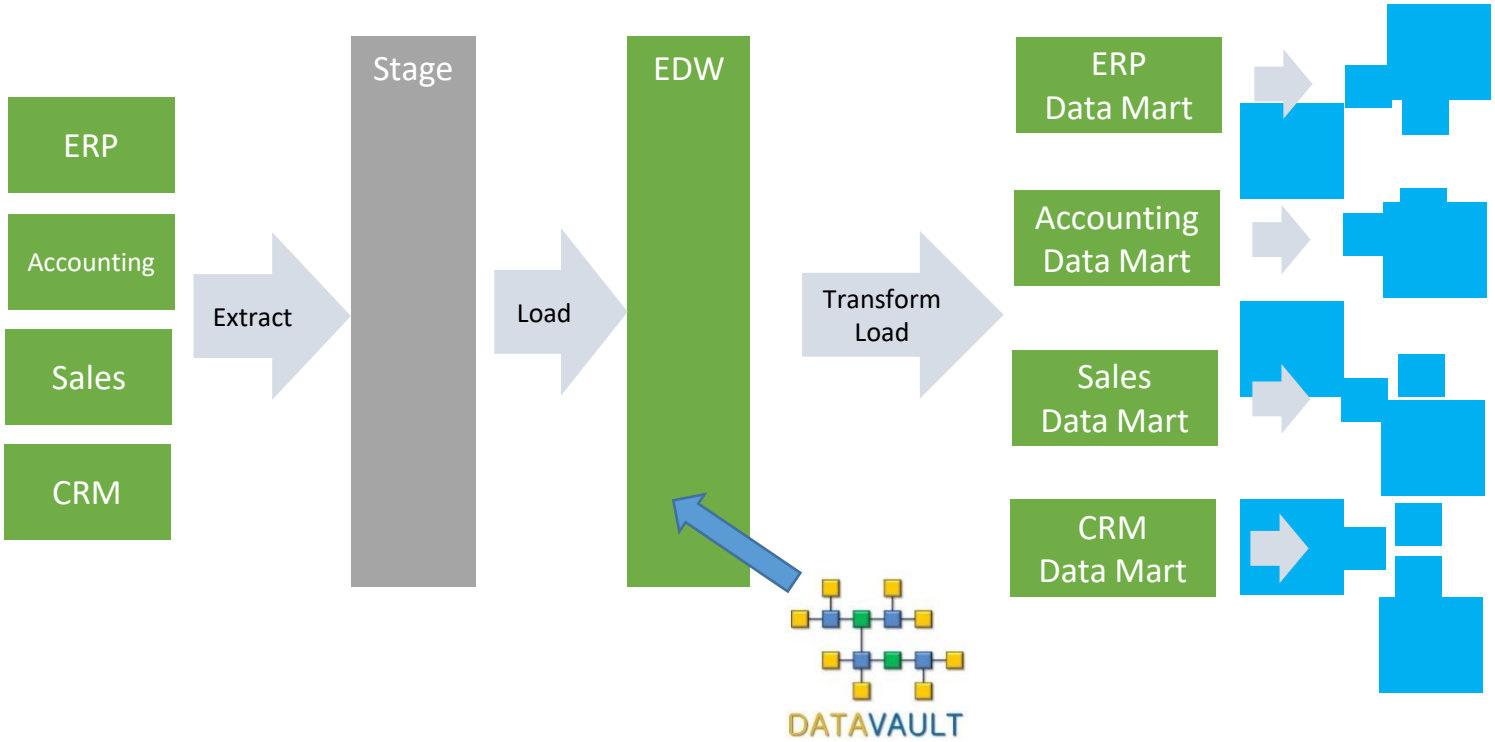
Was ist ein Ensemble? Gibt's das nur bei Data Vault?

- Methodologie (KEINE Technologie!)
- Vereinfacht gesagt: Alles wird Type 2 und sinnvoll gruppiert
- Besteht aus:
  - HUBs
  - LINKs
  - SAT(elitten)S
- HUBs + Links formen den Backbone, Satelliten bringen den Inhalt

Was ist ein Ensemble? Gibt‘s das nur bei Data Vault?

- Verschiedene Ansätze (z.B. Dan Linstedt vs. Hans Hultgren)
  - „Immer“ ist also nicht immer „immer“
  - Es gibt klar falsche aber nie einzig richtige Wege ☺
- Modellierung nach Sprache und Logik des Business
- Eine einmal angelegte Tabelle wird nicht mehr geändert
- Es gibt nur Inserts (Ausnahmen bestätigen die Regel)

Was ist Data Vault? Wo steht das dann inhaltlich?



Was ist Data Vault? HUBs

- Formt den Business Key
- Bestehen „immer“ aus 4 Spalten:
  - Primary Key (Surrogate Key)
  - Business Key
  - Load Timestamp
  - Source System
- Wird immer als erstes befüllt

Was ist Data Vault? Links

- Formt Zusammenhänge / Relationships
- Enthält 5-n Spalten:
  - Surrogate Key
  - Load Timestamp
  - Source System
  - 2-n Foreign Keys (Surrogate Keys der zugehörigen Hubs)

Was ist Data Vault? Satelliten

- Formt Context, Beschreibung und Historie
- Enthält 5-n Spalten:
  - Surrogate Key
  - Foreign Key (Surrogate Key des zugehörigen Hubs)
  - Load Timestamp
  - Source System
  - 1-n Beschreibungsspalten

Was ist Data Vault? Satelliten

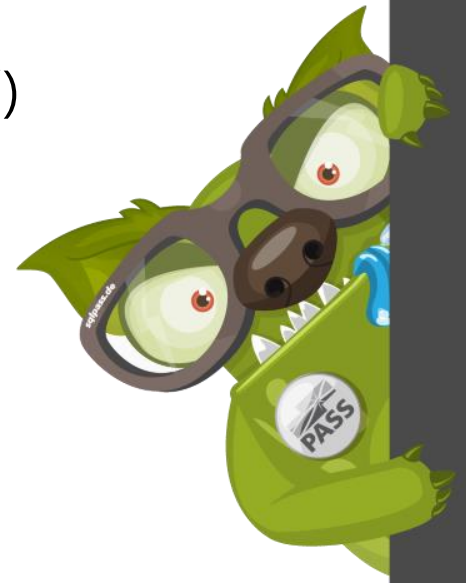
- Ein Hub kann beliebig viele Satelliten haben!
- Daten die nichts miteinander zu tun haben oder sich unterschiedlich häufig ändern gehören also meist in unterschiedliche Satelliten
- Je Business Key ist i.d.R. nur ein Eintrag je Satellit und Zeitpunkt gültig
  - Verschiedene Telefonnummern wären z.B. eher eigene Satelliten mit Typ-Link

Was ist Data Vault? Hub, Link oder Satellit?

- Wozu passt die Überschrift am besten?
  - Kunde
  - Mitarbeiter
  - Zuständiger Mitarbeiter
  - Name des Mitarbeiters
  - Verkaufsmenge in einer bestimmten Transaktion

Was ist Data Vault? Satelliten

- Wie könnte bei den folgenden Kunden-Informationen ein Satelliten Design aussehen?
- (Remember: Es gibt kein eindeutiges „Richtig“)
  - Vorname
  - Nachname
  - Geschlecht
  - Telefonnummer
  - Geburtsdatum
  - Familienstand
  - Adresse
  - Kinder (jeweils Name & Geburtsdatum)



Ist das wirklich so dogmatisch zu sehen?

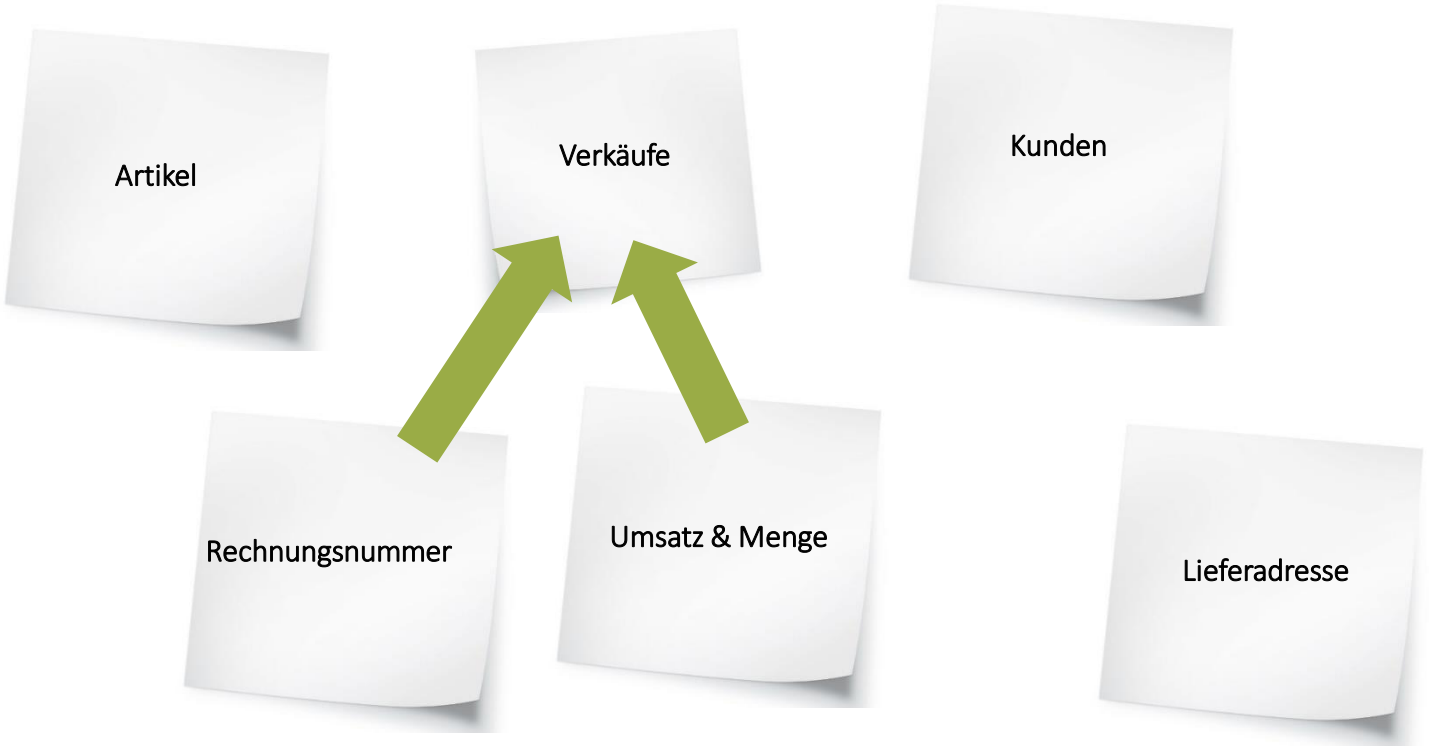
- Für die heutige Session schon ☺
- Generell sind aber „Keyed Instances“, „Linked Sats“, „Valid From/To“, „Multiactive“ etc. durchaus denkbar und erlaubt
- Da kommen dann auch UPDATEs ins Spiel
- (erlaubt ist eh alles, ist ja EUER Modell!)

Wie fängt man da am besten an? Roses are red…

- Mit Post Its!
  - Hubs: Blau
  - Links: Rot
  - Satelliten: Gelb
- Was sind die „Core Business Concepts“ (→ Hubs)?
  - Was beschäftigt das Business? Was sind verständliche Business Keys?
- Wie hängen diese zusammen? (→ Links)
- Was benötigte ich an Kontext? (→ Satelliten)
- Wie sind die Hubs ggf. „aufzubrechen“ aus technischer Sicht?

Wie fängt man da am besten an? Themensammlung

Wie fängt man da am besten an? Themensammlung



Wie fängt man da am besten an? Themensammlung

Artikel

Verkäufe

Umsatz & Menge
Rechnungsnummer

Kunden

Lieferadresse

Wie fängt man da am besten an? Themensammlung

Wie fängt man da am besten an? Themensammlung

Wie fängt man da am besten an? Themensammlung

Artikel

Kunden

Merkerliste

Umsatz & Menge

Rechnungsnummer

Verkaufskopf

Verkaufszeile

Lieferadresse

Wie hängt alles zusammen? Fokus auf HUBs & LINKs

Artikel

Kunden

Merkerliste

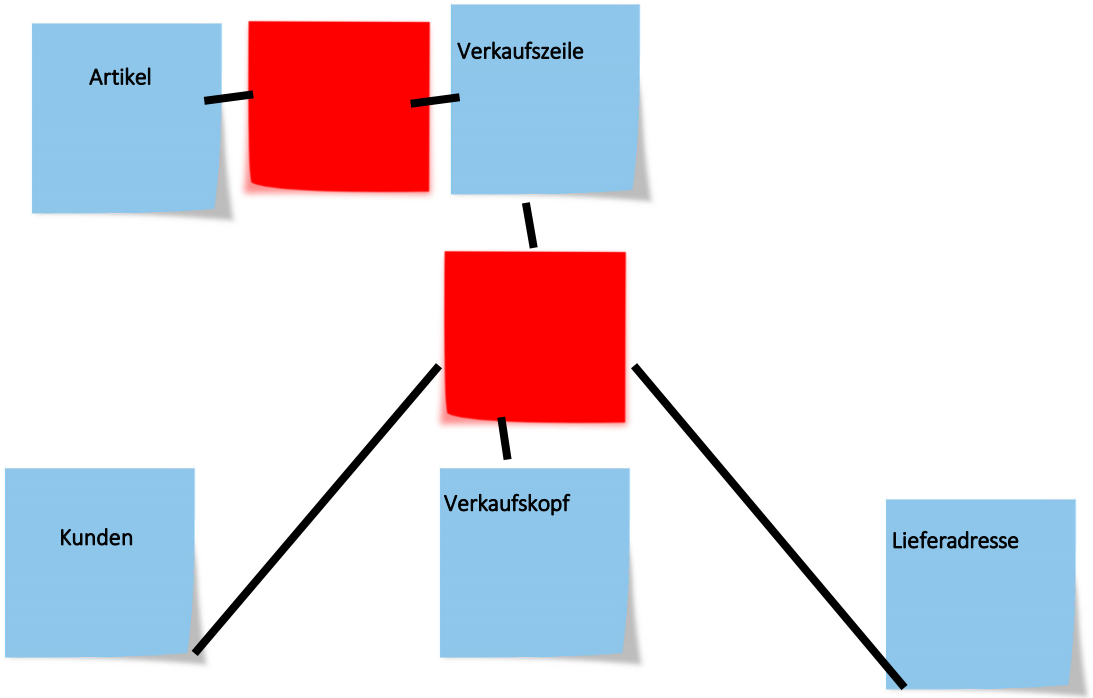
Umsatz & Menge

Rechnungsnummer

Verkaufskopf

Verkaufszeile

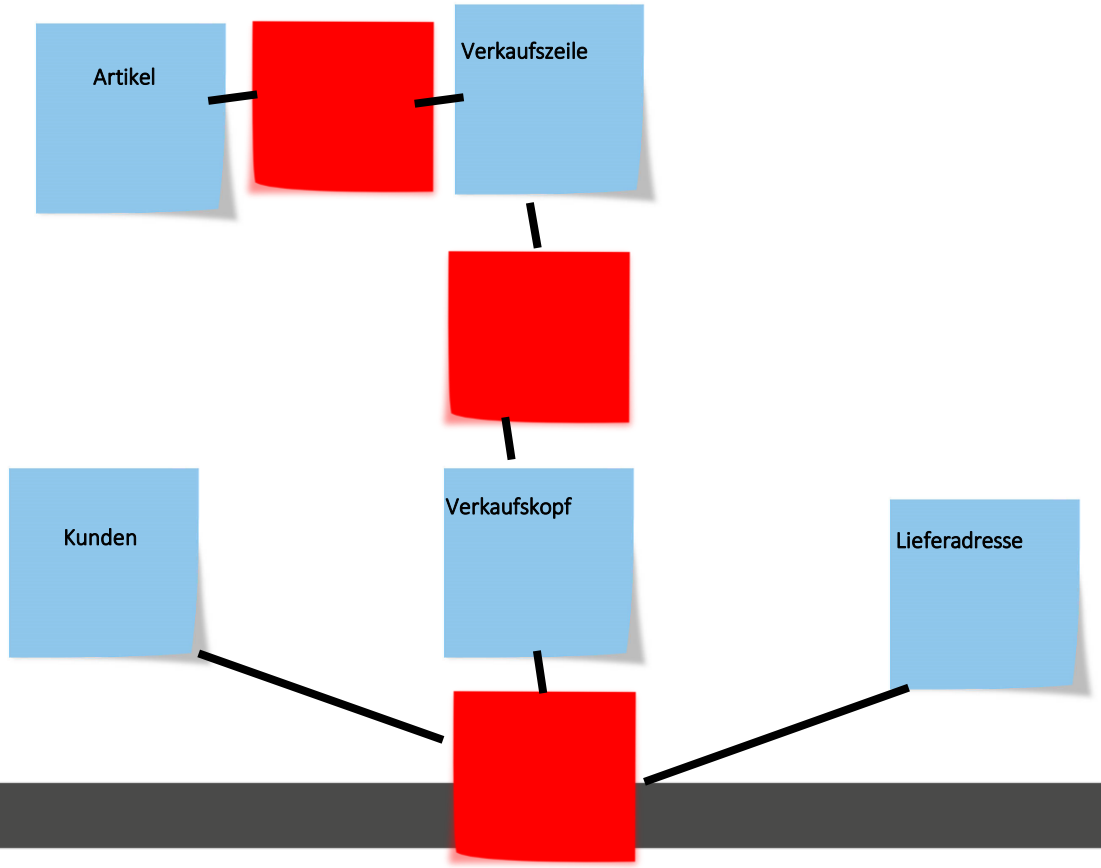
Lieferadresse



Merkerliste

Umsatz & Menge

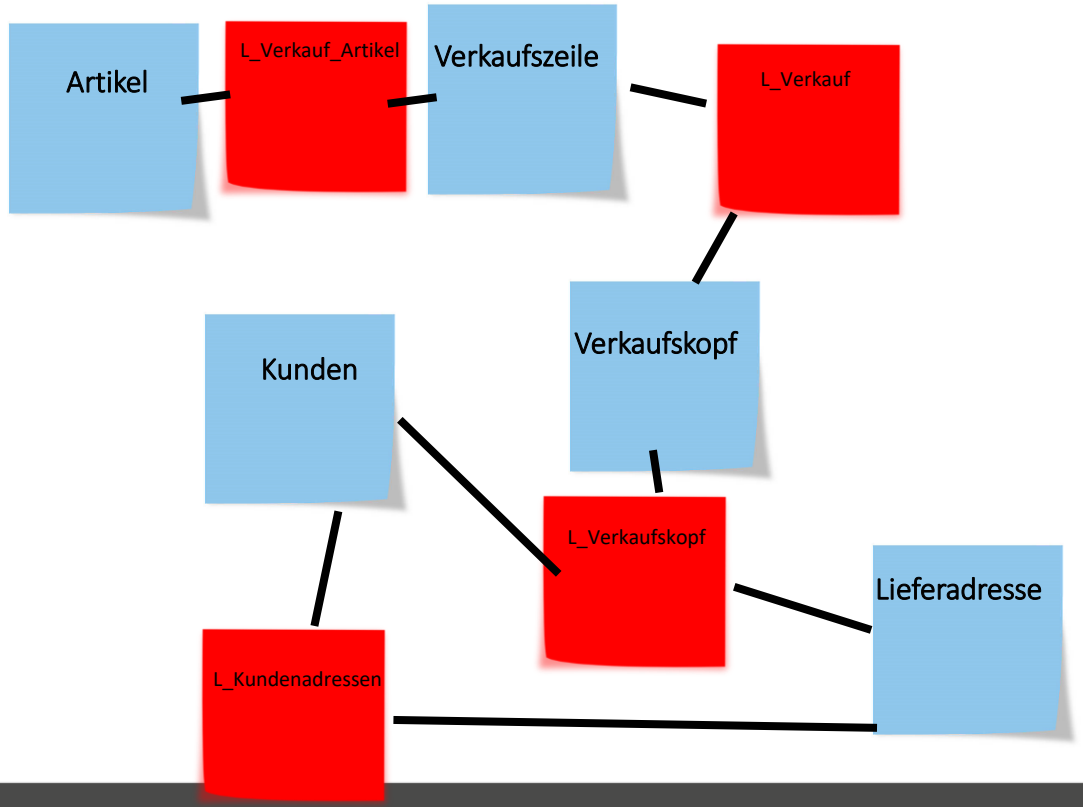
Rechnungsnummer



Merkerliste

Umsatz & Menge

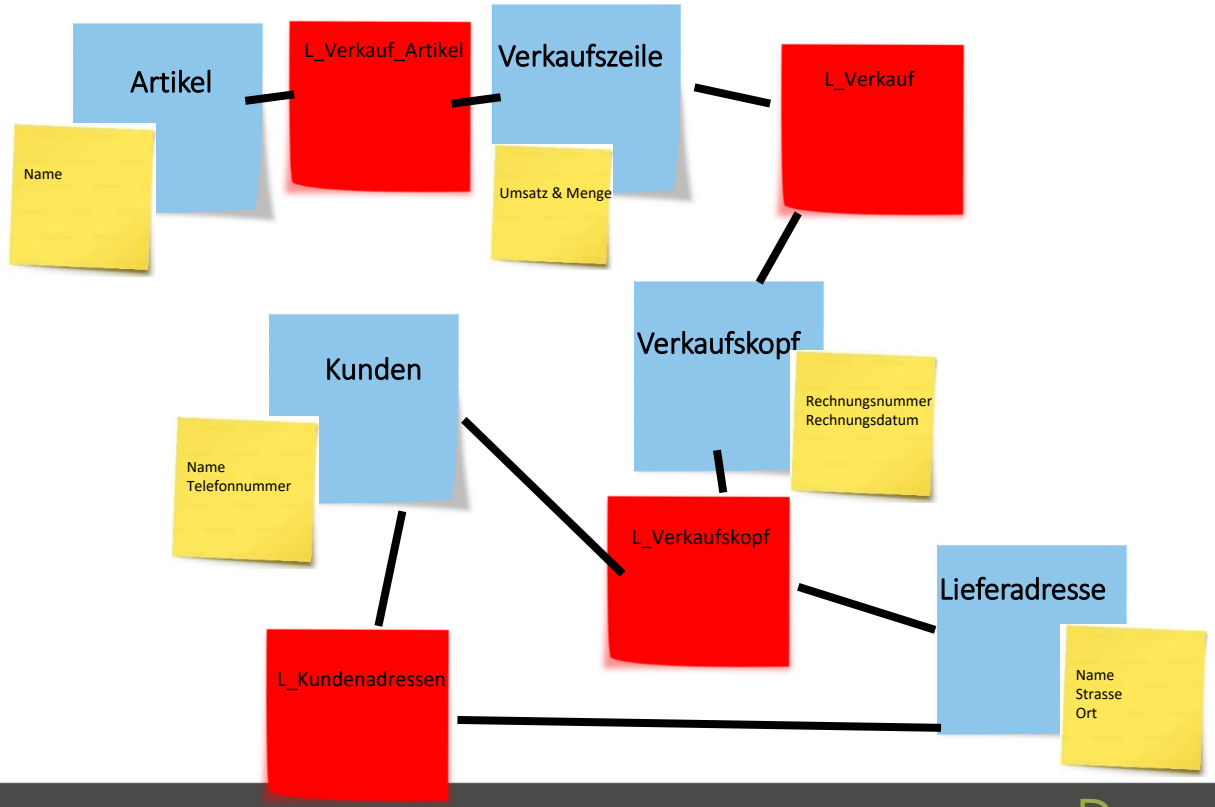
Rechnungsnummer



Merkerliste

Umsatz & Menge

Rechnungsnummer

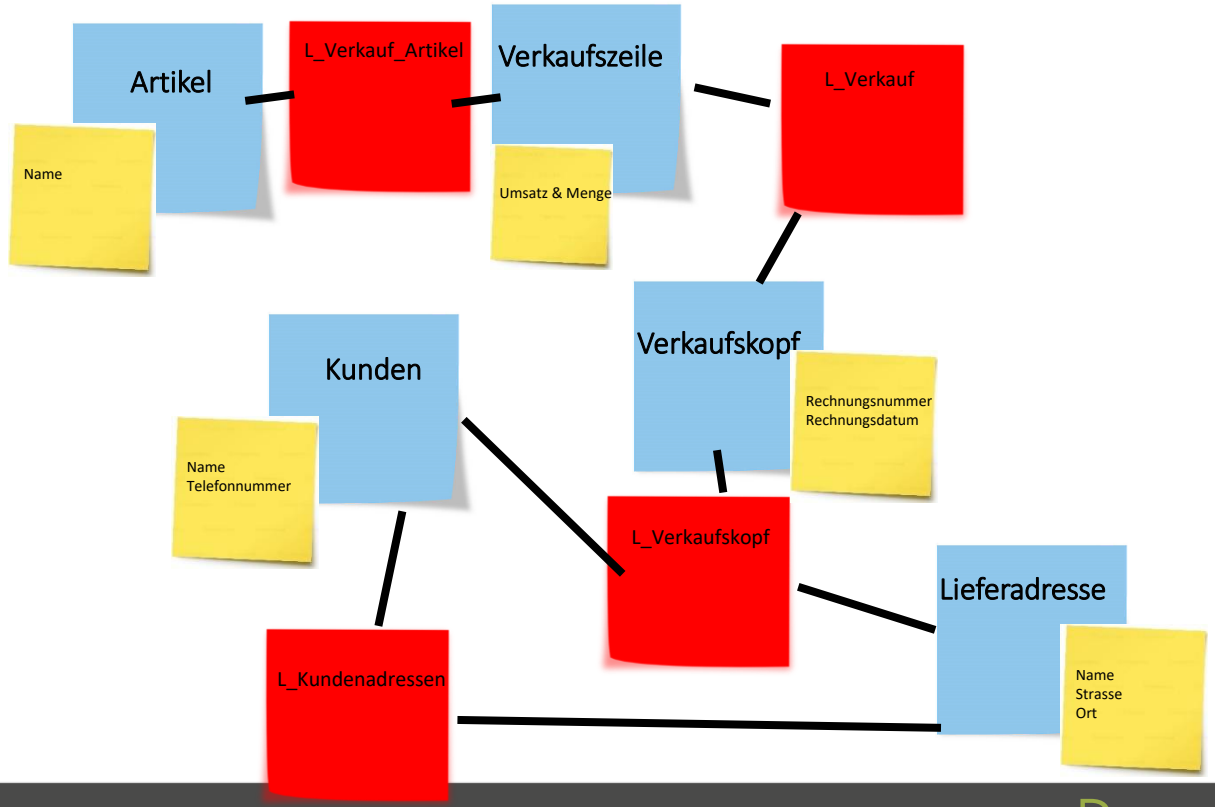


Merkerliste

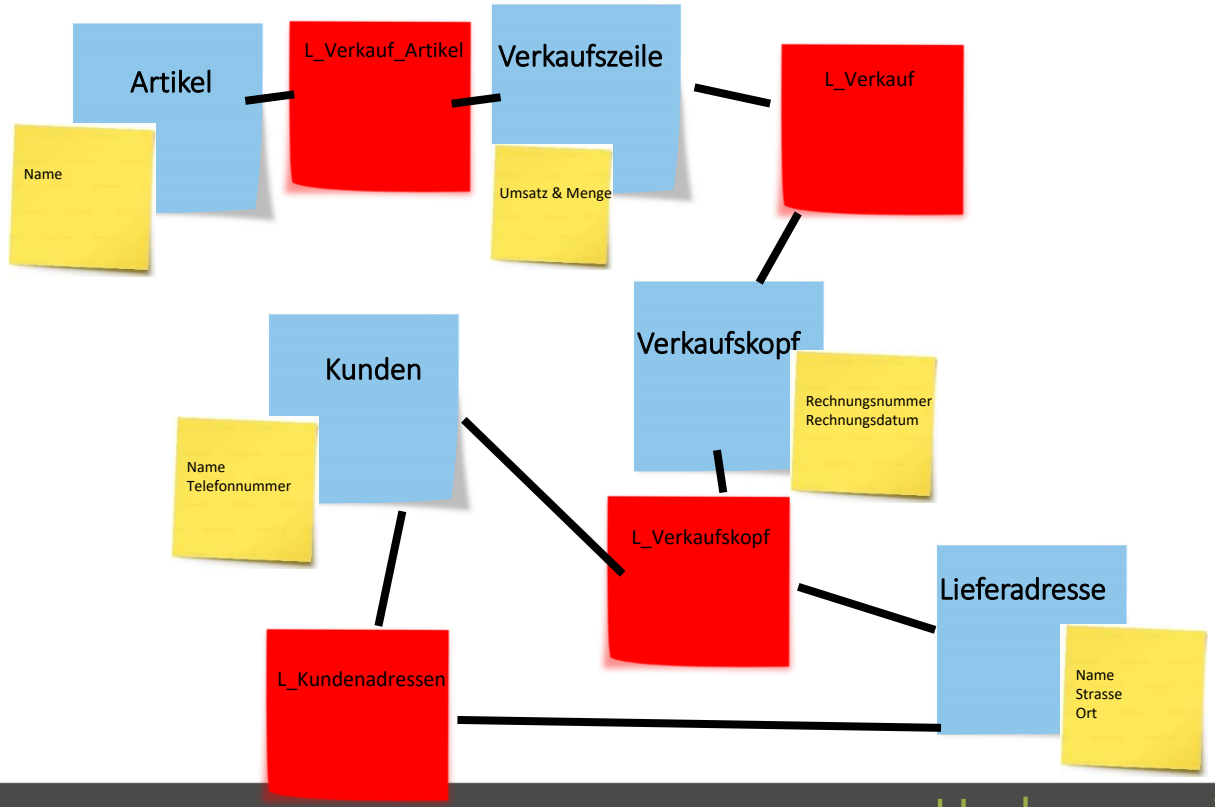
Umsatz & Menge

Rechnungsnummer

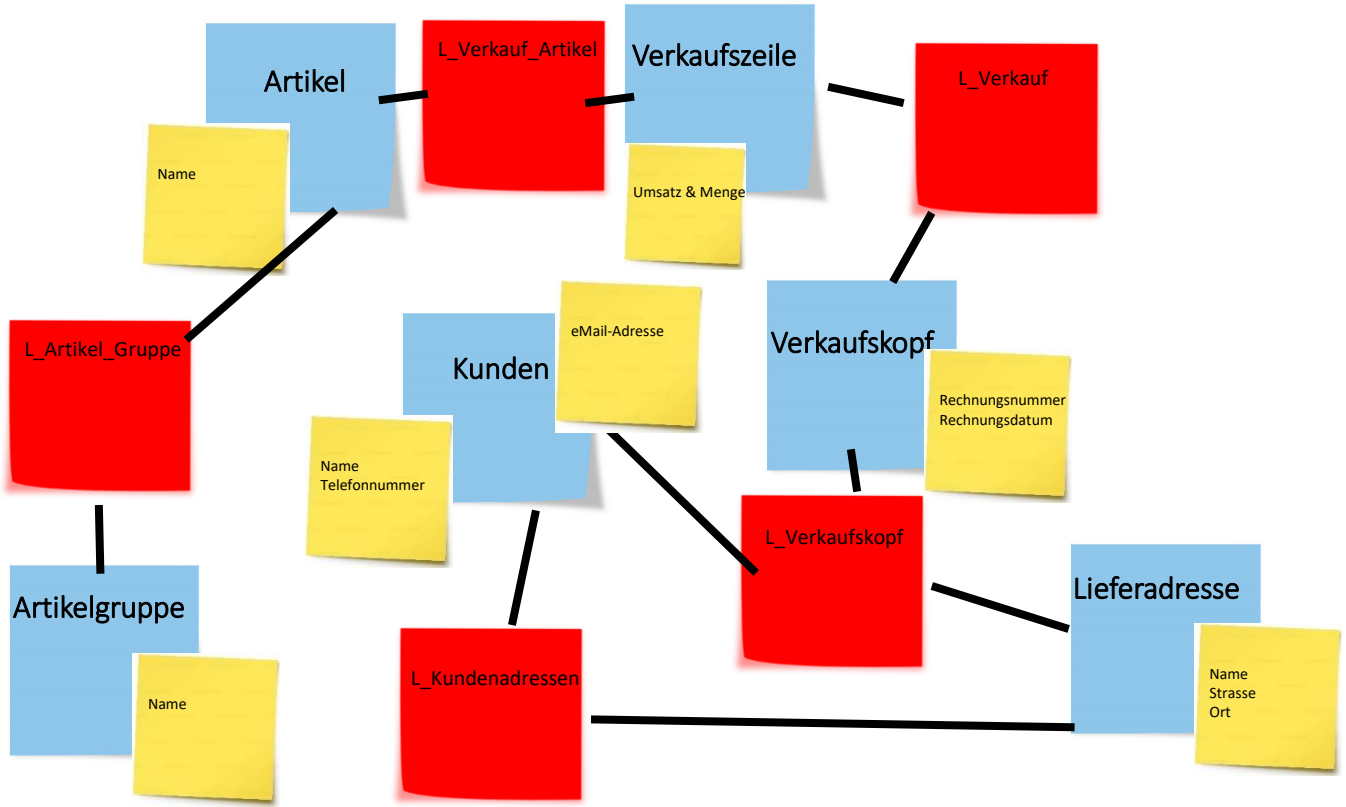
Dann erst die Satelliten



Dann erst die Satelliten



Und wenn sich etwas ändert?



Und wenn sich etwas ändert?

Warum will ich das?

- Nur INSERT Operationen, somit sehr gute Lade-Performance
- Keine Anpassungen an bestehenden Tabellen, somit nicht nur incremental Load sondern auch incremental Build – in einem agilen, iterativen und gut zu automatisierenden Prozess
- Historie für alle Datensätze lösen Auditing und auch eventuelle spätere Point in Time Fragen
- Geschicktes Satelliten Design hilft bei Datenschutz-Compliance

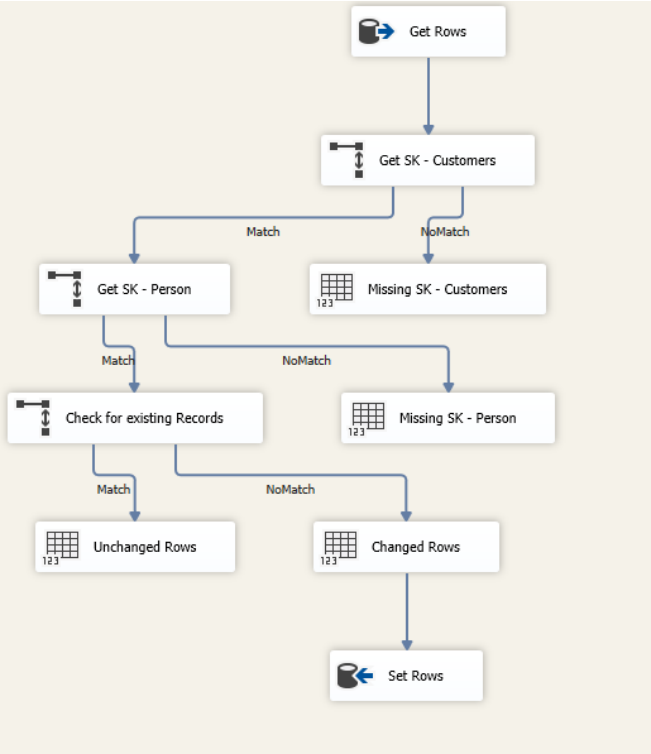
Warum will ich das nicht?

- Viele Extra Joins bringen auch reichlich extra Abfragekosten
  - (Im Gegenzug kann ich aber meine Datamarts sehr schön inkrementell beladen!)
- Erfordert deutlich mehr Wissen über die Struktur des EDW um sinnvolle Abfragen zu erstellen, somit sind für ad hoc Abfragen häufiger Views erforderlich
- Je nach (Vor-)Architektur habe ich sogar zwei Datawarehouses

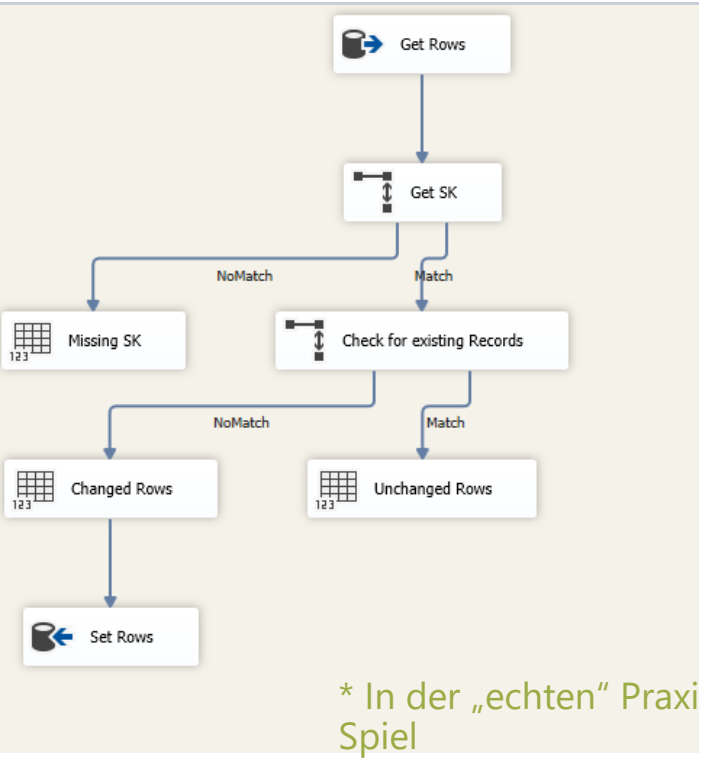
Wie kann das im Praxisfall aussehen? Laden von Hubs*

```
INSERT INTO dv.[HUB_Customers] (H_Customers_BK,Source_Connection)
SELECT BusinessKey,SRC.Source_Connection FROM stage.[AW_Sales_Customer] SRC
LEFT JOIN dv.[HUB_Customers] HUB on SRC.BusinessKey = HUB.[H_Customers_BK]
WHERE HUB.Load_TS IS NULL
```

Wie kann das im Praxisfall aussehen? Laden von Links*



Wie kann das im Praxisfall aussehen? Laden von Satelliten*



* In der „echten“ Praxis kommen hier schnell Hashes ins Spiel

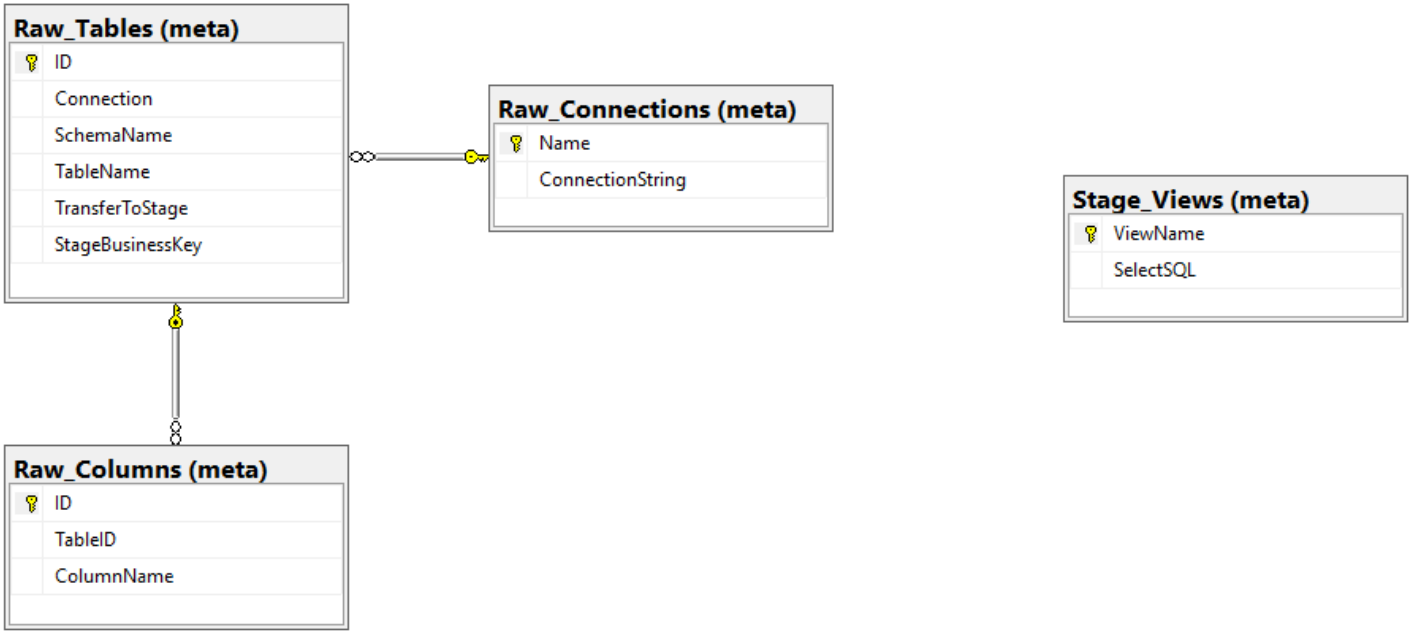
Was ist dieses Hashing?!

- Verschiedenste Methoden/Algorithmen (zum Beispiel MD5)
- Macht aus einem „Stück Daten“ (egal wie groß) ein relativ kleines Stück Daten (bei MD5 zum Beispiel immer 32 Zeichen)
- Vorteil: Nur 1 Lookup
- Problem: Hash Collision (2 verschiedene Inputs geben den selben Output), reduzierbar durch double-hashing Algorithmen
- Beispiel, SQL Hash:
  - SELECT HASHBYTES('SHA1', 'Kundennummer') → 0x1A9DA22C86C6422F6D3C4EBE8195DF40691921BA
  - SELECT HASHBYTES('MD5', 'Kundennummer') → 0x21ADE1995B8E44CF74402025C641C82F
  - SELECT HASHBYTES('MD5', 'Artikelnummer') → 0xDA0DE398714C962BF35EEA7B6ACB6358

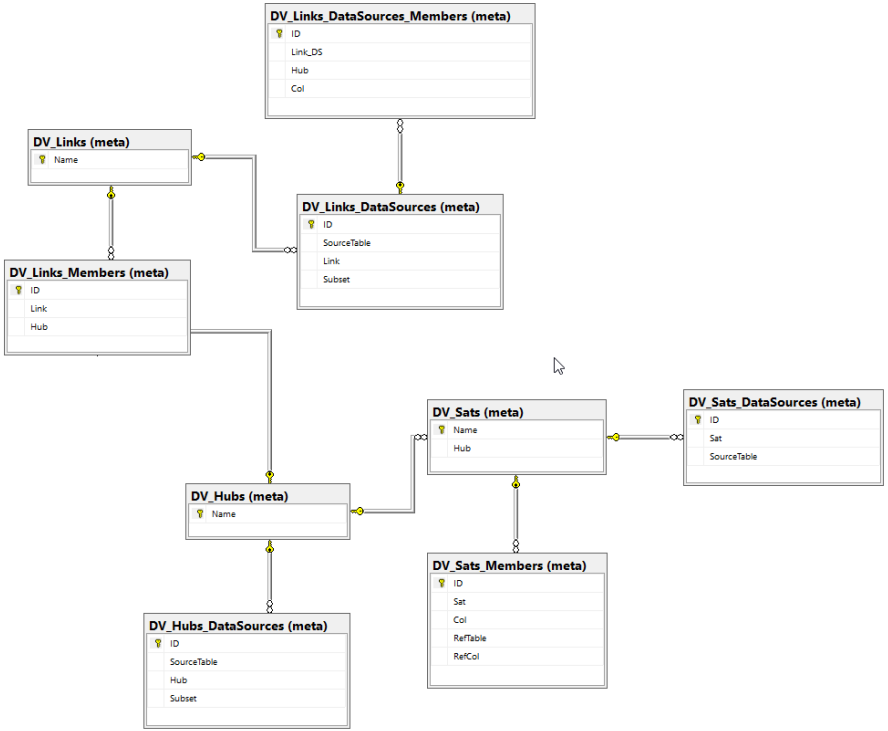
Wie kann das in der Praxis überhaupt funktionieren?

- Nur mit Metadaten
- Nur mit Automatisierung
- Zum Beispiel mit: Biml ☺

Metadaten Modell: Rawstage & Stage



Metadaten Modell: DataVault



Kleines Intro in Biml

- Biml ist eine Markup Sprache – also: XML für BI (SSIS/SSAS/T-SQL)
- Erfunden/Entwickelt von Varigence
- So sieht Biml aus:

```
<Biml xmlns="http://schemas.varigence.com/biml.xsd">
    <Packages>
        <Package Name="HelloBiml"/>
    </Packages>
</Biml>
```

- Wird erst in Verbindung mit BimlScript/APIs mächtig
- Verschiedenste Frontends

Biml Frontends

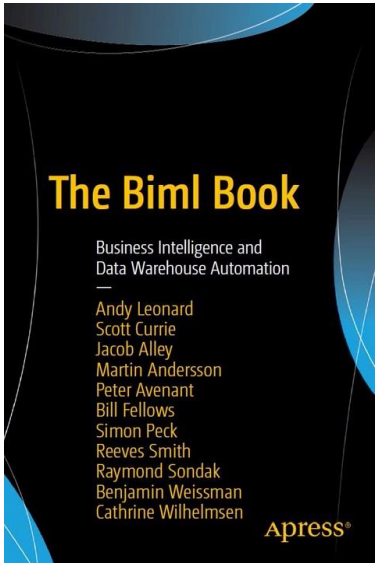
Aha...

Soweit, so gut...
Demo Time

Mehr?

- Data Vault:
  - http://www.GeneseeAcademy.com
  - http://DataVaultBook.blogspot.com/

- Biml:
  - http://biml.blog/
  - http://www.bimlscript.com/
  - http://sqlblog.com/blogs/andy_leonard/archive/2016/06/02/so-you-want-to-learn-more-about-biml.aspx
  - http://www.cathrinewilhelmsen.net/biml/



Fragen?

Gerne auch im Nachgang

per Mail (b.weissman@solisyon.de)

oder

Twitter (@bweissman)

Thank you very much for your attention.
Vielen Dank für Eure Aufmerksamkeit.