Widerspenstige Dimensionen

Dr. Andrea Kennel
InfoPunkt Kennel GmbH
8600 Dübendorf, Schweiz

Schlüsselworte
DWH, BI, Dimensionale Modellierung, Dimensionen.

Einleitung

In einem Data Warehouse werden Marts in Form von Fakten und Dimensionen modelliert. Dabei ist es wichtig, dass für die Auswertungen einheitliche, also „conformed" Dimensionen definiert werden, die mit unterschiedlichen Fakten verknüpft werden können [1].
Dimensionen in „Schulbüchern" sind meist einfach und klar definiert. In der Realität gibt es aber Dimensionen, die sich widerspenstig zeigen, denn sie können komplex sein und vor allem recht unterschiedliche Attribute aufweisen. Zur Zähmung dieser widerspenstigen Dimensionen braucht es die richtige Strategie.
Was kann getan werden, wenn ein Geschäft diverse Produkte verkauft, die auf den ersten Blick komplett unterschiedliche Attribute aufweisen, ab einer gewissen Verdichtung aber doch gemeinsame Attribute haben, die in den Auswertungen interessieren?
Der vorliegende Artikel zeigt die Problemstellung und mögliche Lösungen.

Die Unterschiede in den Produkten

Ein Modegeschäft verkauft Massanzüge, Jeans, Krawatten und Uhren für Damen und Herren. Für die Planung möchte das Geschäft die Verkäufe nach diversen Kriterien auswerten, so beispielweise wie viel für Damenartikel und wie viel für Herrenartikel ausgegeben wird, welche Farben bevorzugt werden und welche Preissegmente wie gut laufen. Weiter möchte aber die Uhrenabteilung zusätzliche Detailauswertungen spezifisch für Uhren. Da interessiert, welche Gehäuse beliebt sind oder welche Armbänder. So wurden die vier Produkte, die verkauft werden, genauer analysiert und als Dimensionen in ADAPT abgebildet.

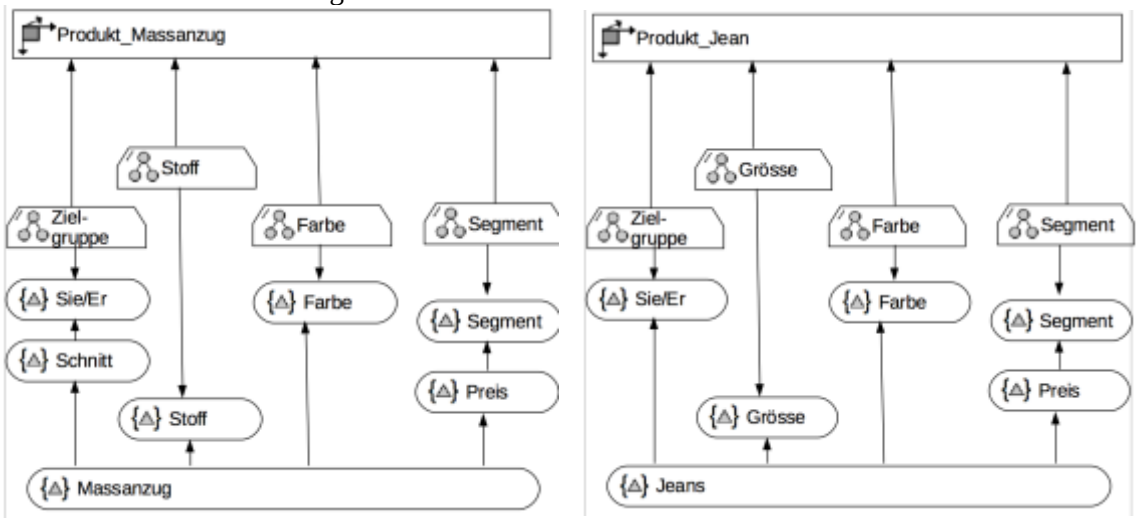

Abb. 1: Die Dimensionen Massanzug und Jeans

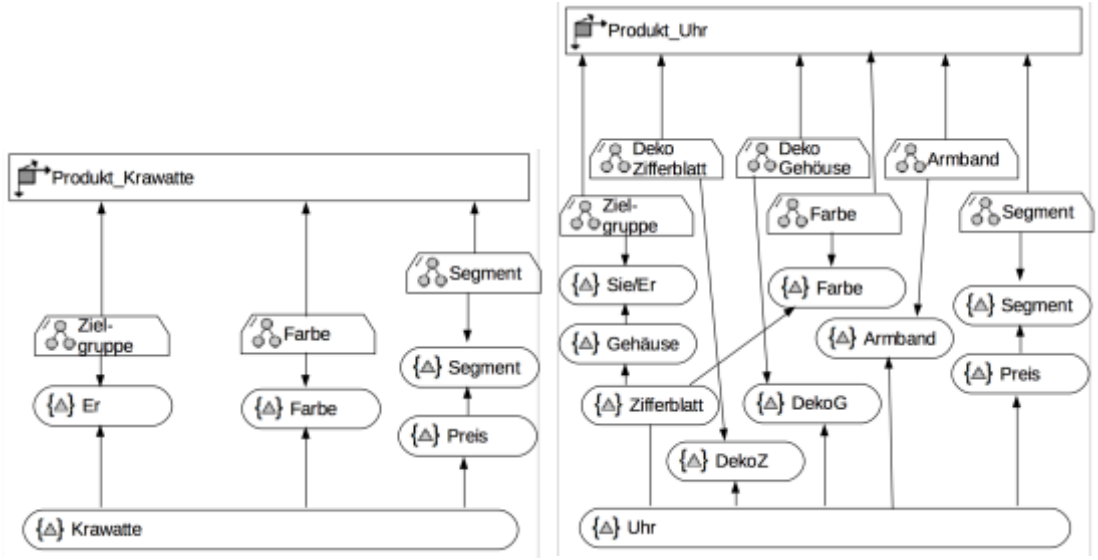

Abb. 2: Die Dimension Krawatte und Uhr

ADAPT [2] zeigt die Dimensionen mit den unterschiedlichen Hierarchien und Hierarchiestufen. Die Begriffe wie Uhr, DekoG etc. bezeichnen nicht die Attribute, sondern die Hierarchiestufen respektive die Hierarchie selber. So hat die Dimension Produkt_Krawatte die drei Hierarchien „Zielgruppe“, „Farbe“ und „Segment“. Die Hierarchie „Segment“ hat auf der obersten Stufe das „Segment“ selber, das das Preissegment beschreibt. Darunter sind die eigentlichen Preise und je Preis gibt es verschiedene Krawatten.

Jeans und Massanzug sind von der Struktur her sehr ähnlich. Krawatte ist das einfachste Produkt und Uhr das komplexeste. Die Uhren haben eine relativ komplexe Struktur, da diese nach Wunsch angefertigt werden und somit fast jede Uhr ein Unikat ist. Bei den Uhren wird zuerst unterschieden, ob diese eine Herren-Uhr oder eine Damen-Uhr ist. Für beide stehen unterschiedliche Gehäuse zur Verfügung. Für jedes Gehäuse wiederum kann ein Zifferblatt ausgewählt werden. Das Zifferblatt legt dann die Farbe der Uhr fest. Sowohl das Zifferblatt als auch das Gehäuse kann mit Edelsteinen und oder Gravuren dekoriert werden. Weiter kann unter verschiedensten Armbädern ausgewählt werden. Der Preis der Uhr variiert sehr stark und hängt von den gewählten Komponenten ab. Je nach Preis wird die Uhr einem Preissegment zugeteilt.

Auch bei den Massanzügen kann der Kunde oder die Kundin verschiedene Ausprägungen wählen. Hier gibt es aber klar weniger Variationsmöglichkeiten als bei den Uhren.

Die Gemeinsamkeiten in den Produkten

Konzentrieren wir uns zuerst auf die Gemeinsamkeiten in den Dimensionen. Gemeinsam sind in allen Dimensionen die Hierarchien „Zielgruppe“, „Farbe“ und „Segment“. Die Hierarchie „Segment“ hat bei allen 4 Dimensionen auch dieselben Hierarchiestufen, ist somit identisch. Die Hierarchie „Zielgruppe“ sieht je nach Produkt anders aus. Bei den Jeans wird nur gerade zwischen „Sie“ und „Er“ unterschieden, bei Krawatten gibt es nur „Er“ und bei Massanzügen und Uhren gibt es weiter Hierarchiestufen. Somit hat die Hierarchie „Zielgruppe“ nicht bei allen Hierarchien dieselbe Granularität, die oberste Hierarchiestufe ist aber bei allen gemeinsam. Ähnlich verhält es sich mit der Hierarchie „Farbe“. Auch hier ist nur die oberste Hierarchiestufe gemeinsam und hat je nach Produkt eine andere Granularität. Das bedeutet, dass nicht alle Hierarchien und Hierarchiestufen für alle Produkte identisch sind.

Basierend auf dieser Analyse stellt sich nun die Frage, wie die Dimension Produkte als „conformed dimension“ [1] abgebildet werden kann.

Eine einheitliche Dimension
Ein erster Ansatz für eine einheitliche Dimension sind die Gemeinsamkeiten. Diese würde folgendes Bild ergeben:

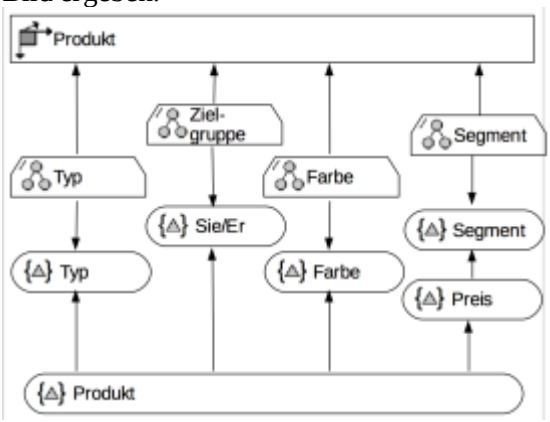

Abb. 3: Dimension mit den gemeinsamen Attributen

Damit die unterschiedlichen Produkte unterschieden werden können, wird die neue, flache Hierarchie Produkttyp eingeführt. So könne die unterschiedlichen Produkte nach den gemeinsamen Eigenschaften ausgewertet werden. Was damit aber nicht möglich ist, sind produktespezifische Auswertungen, wie diese für Uhren gefordert sind.
Damit dies mit der einheitlichen Dimension möglich ist, müssen alle produktespezifische Attribute ergänzt werden.
Diese Dimension hat den Nachteil, dass sie nicht wirklich übersichtlich ist, daher wird hier gar nicht erst versucht, diese grafisch darzustellen. Weiter hat die Dimension auch viele Attribute, die je nach Produkt leer sind. Diese Dimension ist zwar einheitlich, aber dafür widerspenstig und wir sollten prüfen, ob wir diese Dimension bändigen können.

Typisierte Dimensionen
Kinbal beschreibt ein ähnliches Problem [3]. Als Lösung schlägt er typisierte Dimensionen vor. Kurz gesagt werden alle Attribute, die typspezifisch sind, in typisierten Dimensionen zusammengefasst. Die gemeinsamen Attribute und ein Fremdschlüssel auf die typisierte Dimension bleibt in der einheitlichen Dimension.
In vorliegenden Beispiel ergibt das folgendes Bild.

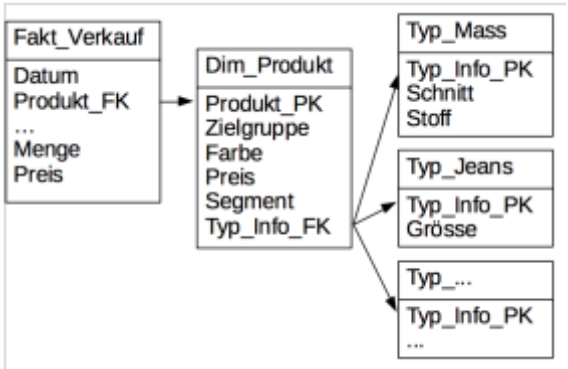

Abb. 4: Produkte als typisierte Dimension

Ein gravierendes Problem bleibt aber mit der Typisierung erhalten. Betrachten wir das Produkt Uhr. Bei einer Uhr können so viele verschiedene Ausprägungen frei gewählt werden, dass kaum zweimal dieselbe Uhr verkauft wird. Das bedeutet, dass für jeden Verkauf einer Uhr ein Eintrag in der Dimension entsteht. Somit erhalten wir eine Dimension mit fast so viele Einträgen, wie es Einträge in der Faktentabelle gibt. Die Aufteilung in eine einheitliche Dimension und in typisierte Dimensionen hilft zwar bei der Bändigung, scheint aber noch nicht die optimale Lösung zu sein.

Minidimensionen

Die Dimension Produkte hat nur 4 gemeinsame Attribute:

Sie/Er	Farbe
Preis	Segment

Werden diese Attribute in die Faktentabelle übernommen, so können diese auch sehr einfach ausgewertet werde. Gleichzeitig wird der Fremdschlüssel auf die Dimension in der Faktentabelle so angepasst, dass dieser direkt auf die typisierte Dimension zeigt. Für die Auswertung über alle Produkte wird keine Dimension mehr gebraucht, da alle Attribute direkt in der Faktentabelle vorhanden sind. Für produkttypenspezifische Auswertungen wird nur die entsprechende typisierte Dimension verknüpft. Dadurch werden auch diese Auswertungen beschleunigt.

Betrachten wir diese Lösung aber genauer, so ist zwar die Dimension nicht mehr widerspenstig, dafür aber die Faktentabelle. Jede Faktentabelle, die mit Produkten zu tun hat, muss dann neben dem Fremdschlüssel auf die Dimension auch die gemeinsamen Attribute enthalten. Ändert sich die Dimension Produkt und erhält beispielsweise eine neue, gemeinsame Hierarchie „Qualität", so muss dieses Attribut in allen Faktentabellen nachgetragen werden. Das ist nicht nur widerspenstig, sondern widerspricht der Idee vom Fakten und Dimensionen.

Junk-Dimension

In einer Junk-Dimension [4] können die wichtigsten, gemeinsamen Attribute zusammengefasst werden. Wir wissen, dass von den 4 gemeinsamen Attribute der Preis nicht relevant ist. So bleiben noch „Sie/Er" mit 2 Werten, „Farbe" mit 12 Werten und „Segment" mit 4 Werten. Das ergibt 96 Kombinationen. Diese können in einer Junk-Dimension zusammengefasst werden. Was geschieht mit dem Preis? Dieser kann als Faktum in der Faktentabelle abgelegt werden oder aber in den typisierten Tabellen, die neben der Junk-Dimension bestehen bleiben.

Die gewünschten Auswertungen für das ganze Geschäft, sprich über alle Produkte, kann einfach und schnell über die Junk-Dimension erstellt werden. Eine Erweiterung dieser Dimension um ein weiteres Attribut wie „Qualität" ist einfach möglich, ohne dass die Fakten-Tabelle oder sogar mehrere Fakten-Tabellen angepasst werden müssten. Gleichzeitig ist aber auch eine Auswertung über einen spezifischen Typ eines Produktes möglich.

Anders betrachtet kann man sagen, dass nur die wirklich einheitlichen Attribute in einer separaten Junk-Dimension zusammengefasst sind. Zusätzlich werden die typspezifischen Attribute in typisierten Dimensionen gespeichert, die für sich auch wieder einfach erweiterbar sind, ohne dass dies auf die Fakten-Tabellen einen Einfluss hätte.

Mit der Kombination von typisierten Dimensionen für die typspezifischen Attribute und einer Junk-Dimension für die gemeinsam relevanten Attribute kann die widerspenstige Produkte-Dimension gezähmt werden.

Referenzen

[1] Ralph Kimball, Margy Ross; The Data Warehouse Toolkit; John Wiley & Sons, 01.07.2013
[2] ADAPT: http://www.infokennel.ch/j3/images/pdfs/DW2016-Referat.pdf
[3] Ralph Kimball, Margy Ross; The Kimball Group Reader; John Wiley & Sons, 2010; Seite 336
[4] Ralph Kimball, Margy Ross; The Kimball Group Reader; John Wiley & Sons, 2010; Seite 306

Kontaktadresse:

Dr. Andrea Kennel
InfoPunkt Kennel GmbH
Am Wasser 3
CH-8600 Dübendorf

Telefon: +41 (0) 44-820 71 40
E-Mail andrea@infopunkt.ch
Internet: www.infopunkt.ch