Hochschule für Technik, Wirtschaft und Kultur

Fakultät Informatik, Mathematik und Naturwissenschaften

Projekt zur Lehrveranstaltung Data Warehousing

Betriebswirtschaftliche Analyse und Entwurf des Data Warehouse

Datum: 2. Juli 2017

Verfasser: Matthias Zober und Enrico Wüstenberg

Inhaltsverzeichnis

1 Phase 1: Analyse **1**

1.1 Beschreibung möglicher Anwendungen aus Business-Sicht 1

1.2 Konzeptuelle Modellierung 3

1.3 Datenverarbeitungsanforderungen 4

2 Phase 2: Entwurf des Data Warehouse **6**

2.1 Relationale Umsetzung eines MDM-Schemas 6

2.2 Optimierung der Data Cubes 7

1 Phase 1: Analyse

1.1 Beschreibung möglicher Anwendungen aus Business-Sicht

Um den längerfristigen Erfolg eines Unternehmen zu gewährleisten, ist es notwendig Kennzahl und Faktoren für dieses zu definieren. Aus den gegebenen Zahlen lassen sich verschiedene Perspektiven ableiten. Die somit definierten Perspektiven sollten finanzielle Aspekte berücksichtigen, das Kaufverhalten des Kunden veranschaulichen und die Entwicklung des Unternehmens allgemein beschreiben.
Im folgenden werden die einzelnen Perspektiven: Finanzperspektive, Kundenperspektive und Entwicklungsperspektive vorgestellt.

Finanzperspektive

Ziel oder Fragestellung	**Kennzahlen**
Wieviel kauft das Bundesland? Kaufkraft der Bundesländer im Verhältnis zur Einwohnerzahl	Vergleich Verkaufsanzahl zu Einwohnerzahl (Quantität, zeitlich, ortsbezogen)
Wie wirkt sich die Farbe eines Artikels auf den Preis und den verkauften Einheiten aus?	Artikeleigenschaft (Farbe), Preis und Verkaufsanzahl
Wie hoch ist die Marge einer Artikelgruppe? Preise von Artikelgruppen im zshg. mit dem Markwert von Rohstoffen	Marktwert der Rohstoffe und Preis von ausgewählten Artikelgruppen (z.B.: Shirt)

Tabelle 1: Finanzperspektive

Kundenperspektive

Ziel oder Fragestellung	**Kennzahlen**
Werden die Rechnungen von einer anderen Person gezahlt? Welcher Altersgruppe gehört der zahlende an? Analyse des Kaufverhaltens	Eigenschaften von Kunde mit Bestellung und zahlenden Kunden
Zshg. von Körpergröße mit online/offline Einkäufen und Retourenanzahl	Eigenschaften Artikel (Größe), Eigenschaften Bestellung (online), und Anzahl der Retouren

Tabelle 2: Kundenperspektive

Entwicklungsperspektive

Ziel oder Fragestellung	Kennzahlen
Welche Artikelgruppe (z.B.: Shirt) wird wann in welchen Bundesland verkauft und wie teuer wäre die Herstellung?	Verkauf von ausgewählten Artikelgruppen pro Land und Zeit mit Rohstoffpreis
Wie zahlen die Generationen?	Umsatz nach Alter pro Jahr
Wer kauft online o. offline, nach Altersgruppe und Bundesland	Eigenschaften Bestellung, Ort und Alter des Kunden

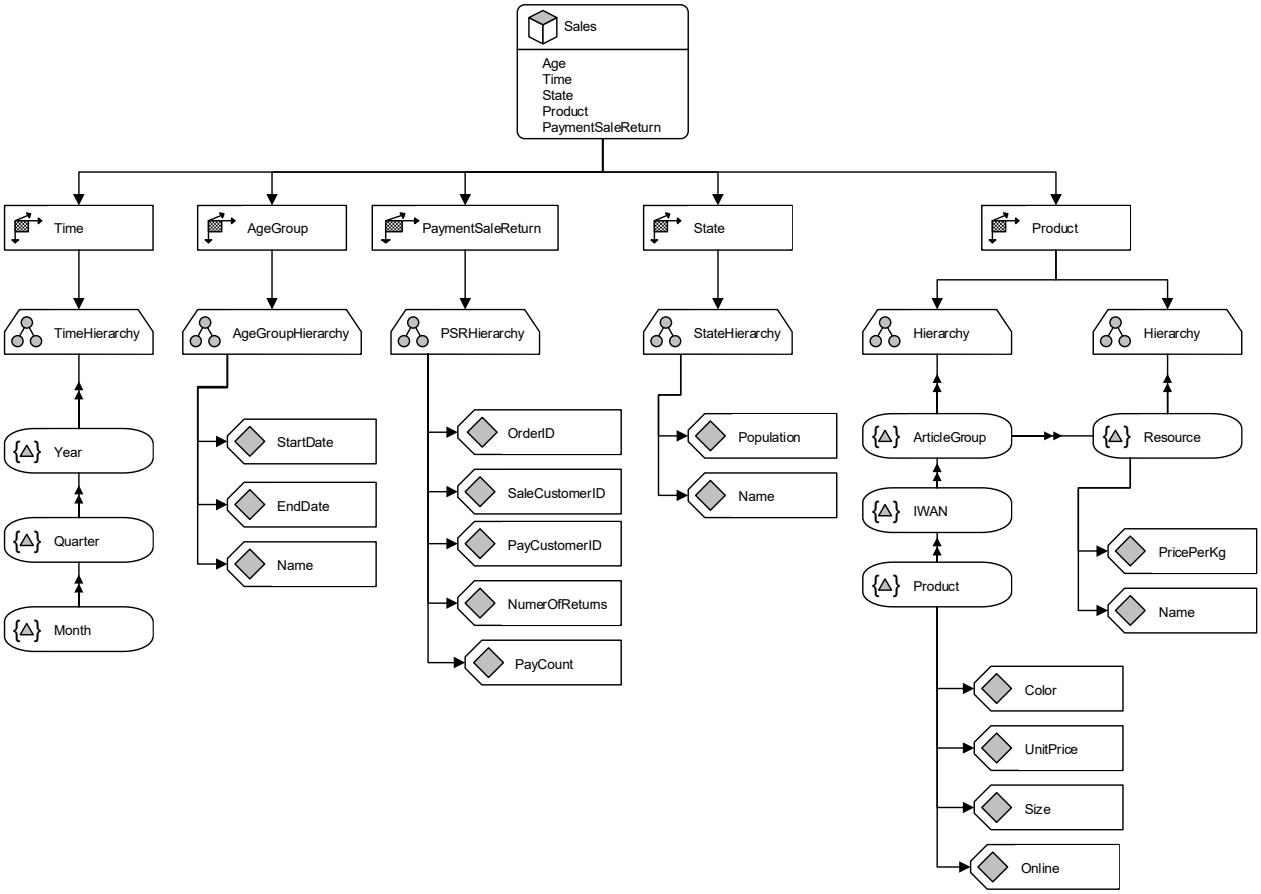
Tabelle 3: Entwicklungsperspektive

Besonderheiten der Zielstellungen

Die gestellten Ziele und Fragestellungen beziehen sich nicht auf übliche Auswertungen, wie etwa die Lebendsdauer eines Kunden oder Warenkorbanalysen. Die gegebenen Kennzahlen wurden so gewählt, dass möglichst interessante Fragestellungen trotz des Verzichts behandelter Abfragen der Seminarserien entstanden. Für die Umsetzung dieser Analysen muss neben der Arbeit auf den gegebenen Daten ein Mehraufwand getätigt werden. Zum einen müssen für die Auswertungen externe Quellen verwendet werden, wie etwa für die Rohstoffe oder die Zuordnung der Postleitzahl zu den Bundesländern. Zum anderen müssen selbst definierte Gruppierungen gebildet werden, wie etwa die Artikelgruppe ”Shirts”.

1.2 Konzeptuelle Modellierung

Cube Bestellungen



Der dargestellte Cube soll für alle angeforderten Auswertungen verwendet werden. Neben den Standard-Hierarchien wie Zeit, Altersgruppe oder Ort (`State`) gibt es auch die Produktdimension und die Dimension der abgeschlossenen Bestellungen (`PaymentSaleReturn`). Produktspezifische Auswertungen wie die Eigenschaften eines Artikels als auch die Eigenschaften von Artikelgruppen, wie die verbrauchten Ressourcen, können mit der Produktdimension getätigt werden. Alle zahlungstechnischen Details einer Bestellung oder Retoure (z.B. Wer zahlt die Rechnung?), werden in der Dimension der `PaymentSaleReturn`-Dimension zur Verfügung gestellt.

1.3 Datenverarbeitungsanforderungen

Finanzperspektive

Im finanziellen Sektor stellen wir uns die Frage, wie die Kaufkraft in Abhängigkeit von der Einwohnerzahl (erhoben 2015) auf alle Bundesländer verteilt ist. Diese Auswertung kann einmal jährlich vorgenommen werden, um eine Tendenz zu beschreiben. Einwohnerzahlen werden jährlich erhoben, folgen aber einer berechenbaren Wachstumsrate, so dass man mit einem festen Wert rechnen kann. Zugleich kann man auch einen monatliche Report ausgeben, doch dieser würde keinen Trend anzeigen.

Bei der Auswahl eines Produktes fällt das Aussehen zu 90% ins Gewicht. Es ist zu prüfen, welches Produkt in welcher Farbe die höchsten Absätze generiert. Bei neuen Produkten kann diese Analyse schon wöchentlich geschehen, damit nicht rentable Produktfarben ausgeschlossen werden können. Diese Aussage hat allerdings keinen Einfluss auf die Qualität eines Produktes oder ob der Erfolg davon abhängig ist. Ein schlechtes Produkt ist auch in einer schönen Farbe schlecht. Man kann nur eine Aussage über die Vorliebe einer Farbe treffen und kann so eine Vorausplanung machen, wie oft das Produkt in der Farbe verkauft wird.

Das dritte Ziel der Finanzanalyse soll eine grobe Übersicht zwischen dem Preis des Rohstoffes und dem Verkaufspreis sein. Die Produkte werden in Artikelgruppen eingeordnet und diesen dann Rohstoffen. Hier werden hauptsächlich Textilien verkauft, daher wurden Rohstoffpreise für Baumwolle, Seide, Leder und Polyester aus dem Jahr 2012 ermittelt. Pro Artikelgruppe wird auch ein durchschnittliches Gewicht berechnet und so der Herstellungspreis bestimmt. Diese Aussage ist nicht sehr genau, da Kosten für Verwaltung, Transport usw. fehlen.

Kundenperspektive

Unser erstes Ziel ist es, in Erfahrung zu bringen wer die Rechnung der Bestellung bezahlt. Hier können Kunden gefiltert werden, die sich ihre Bestellungen bezahlen lassen. Die Schlussfolgerung daraus ist weit interpretierbar. Zum Beispiel kann angenommen werden, dass es sich um ein Geschenk handelt oder ob die Großeltern/Eltern häufig die Rechnungen begleichen. Hier kann nur mit dem Alter eine Aussage getroffen werden, da die Namen anonymisiert wurden. Wird die Rechnung z.B. von jemanden bezahlt der 20 Jahre älter ist, so kann es sich um ein Elternteil handeln. Allerdings ist auch hier

der Report nicht sehr aussagekräftig, da es sich hier auch um Geschwister, Verwandte oder Freunde handeln kann. In Kombination mit dem Nachnamen und der Stadt kann diese Aussage allerdings genauer werden. Gleichzeitig kann ermittelt werden, welche Altersgruppe in dieser Filiale einkauft und welchen Betrag sie monatlich/jährlich da lässt. Diese Angaben sind sehr genau, da Alter und Betrag eines Kunden erfasst werden. Diese Analyse monatlich auszuführen ist überflüssig, da eine gute Aussage erst auf lange Sicht möglich ist.

Unser drittes Ziel ist es, in Erfahrung zu bringen, ob Menschen mit einem großen Körperindex eher Online-Einkäufe tätigen oder ob es keinen Unterschied zu dem lokalen Einkauf gibt. Zugleich kann in Erfahrung gebracht werden, ob sich der Ausbau eines Onlinehandels lohnt. Hier gibt es genaue Zahlen, welche Produkte online gekauft wurden, in welcher Größe und auch wie oft. Es kann also eine zuverlässige Aussage getroffen werden. Doch man muss beachten, dass man mit der Körpergröße nicht gleich auf den Körperumfang schließen kann. Die Retourenanzahl ist zudem ein Faktor, um zu ermitteln, welcher Kunde mehr Kosten als Nutzen verursacht. Auch hier können genaue Zahlen gemacht werden. Dieser Report sollte monatlich ausgewertet werden.

Entwicklungsperspektive

Unser erstes Ziel kann mit der Finanzanalyse verbunden werden und kann darauf Bezug nehmen.

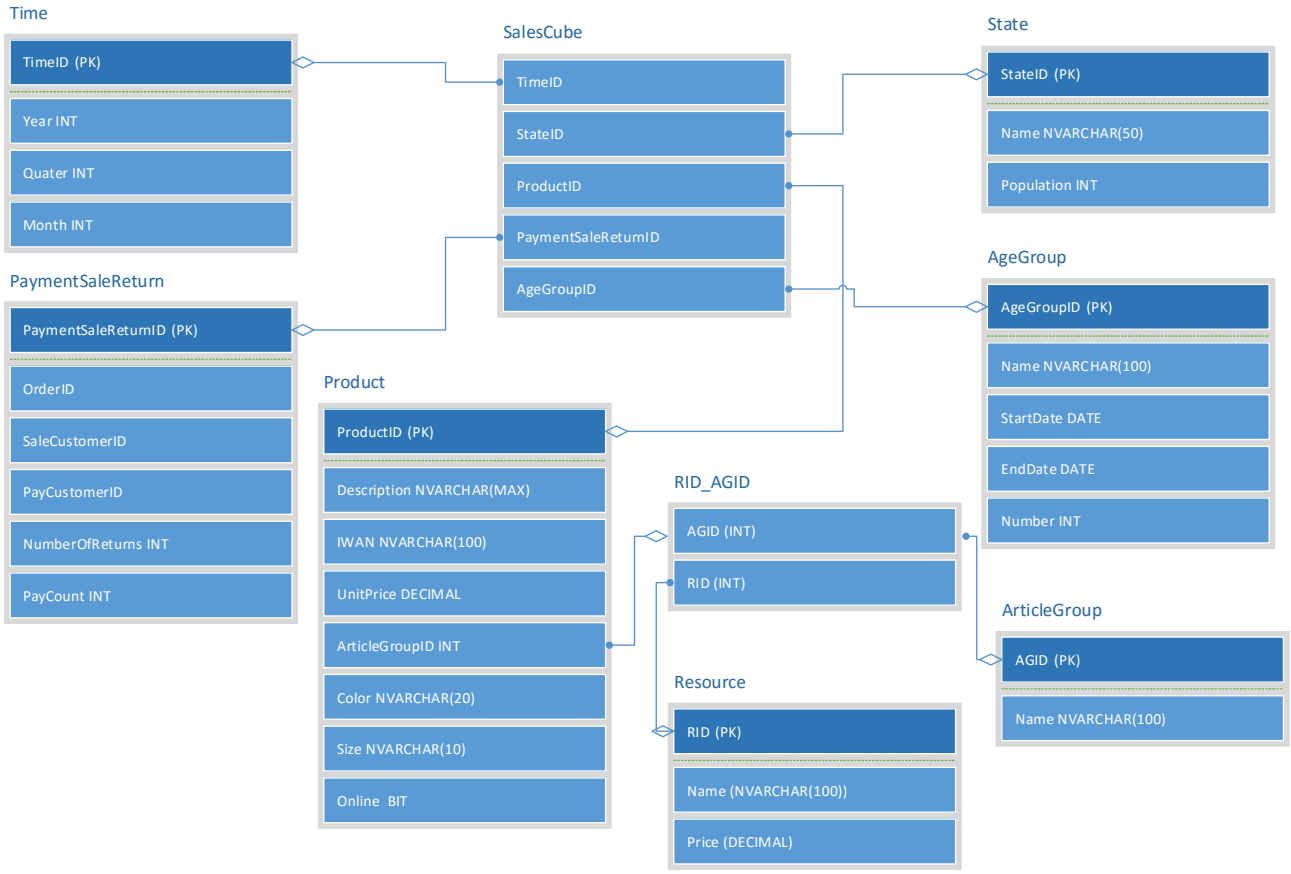
Ein interessanter Fakt ist der Einkauf der verschiedenen Altersgruppen. Zahlen Ältere eher Bar und die Jungen eher mit Karte? Diese Daten sind sehr genau in der Datenbank hinterlegt und können auch wöchentlich ausgewertet werden. Je nach Aktualität der Datenbank können auch Echtzeitanalysen erhoben werden. Hier kann nebenbei untersucht werden, ob z.B. die Terminals einer Filiale ausgefallen sind, weil auf einmal nur noch Bar bezahlt wird. Ursachen können fehlerhafte Geräte oder Einbruch des Banksystem sein. Um eine genau Aussage zu treffen, müssen die Daten in Echtzeit ausgewertet werden und mögliche andere Fehlerquellen (wie z.B. Ausfall der Internetverbindung) berücksichtigt werden.

2 Phase 2: Entwurf des Data Warehouse

2.1 Relationale Umsetzung eines MDM-Schemas

Für die Realisierung, der in Phase 1 entworfene Data Cube wird das nachfolgende relationale Modell verwendet. Hierbei wurde zwischen dem Star- und dem Snowflake-Schema abgewogen. Ein Star-Schema würde zwar die Struktur des Cubes weiter vereinfachen, aber dafür die Redundanz steigern und die Größe der Tabellen ausweiten. Ein Snowflake-Schema sorgt zwar für normalisierte Tabellen ohne Redundanzen, dennoch ist die Anfrageverarbeitung für solche Strukturen schwer und die Performance durch die hohe Anzahl der JOIN-Bedingungen schlechter. Für die Umsetzung des Cubes wurde das Star-Schema ausgewählt. Hierbei gilt die Einschränkung, dass die Cube-Tabelle nur aus Foreign Keys besteht. Diese Einschränkung hat zur Folge, dass die Dimension `PaymentSaleReturn` entstand und die Cube-Tabelle `Sales` nicht all zu groß ist. Diese Entscheidung basiert auf dem Gedanken, dass diese öfters verwendet wird und somit die Performance für die Abfragen besser ist. Das MDM-Schema wurde somit aus Sicht der Performance und der Anfrageverarbeitung optimiert.

Abbildung 1: Relationales Schema des Cubes Bestellungen



2.2 Optimierung der Data Cubes

Neben der Optimierung des Data Cubes aus Schema-Sicht (siehe 2.1), kann er aus Sicht der gestellten Anfragen optimiert werden.

Materialisierte Sichten Eine Performance-Optimierung stellt hierbei die Nutzung von materialisierten Sichten dar, diese können sowohl für den Cube als auch für die Basis-Datenbank erstellt werden und somit den Analyse- als auch den Lade-Prozess beschleunigen. Für diese Sichten wäre in unserem Fall eine monatliche Aktualisierung ausreichend.

Index-Strukturen Eine weitere Optimierung stellen Index-Strukturen dar. Der vorliegende Data Cube kann unter verschiedenen Gesichtspunkten mit Indizes angereichert werden. Zum einen können häufig abgefragte Attribute mit kleiner Domäne mit **Bitmap-Indizes** ausgestattet werden, hierzu zählt z.B. das Attribut "Online" von der Tabelle Dimensionstabelle Produkt. Außerdem können **Bitmap-JOIN-Indizes** bei `JOIN`-Partnern mit niedriger Kardinalität verwendet werden, hierzu zählen z.B. die Attribute der Ressourcen und der Artikelgruppen. Für häufig genutzte Attribute mit großer Domäne bieten sich klassische Indexe aus den relationalen OLTP-Datenbanken wie etwa **B-Tree-Indexe** an. Wenn multidimensionale Abfragen sehr häufig stattfinden ist es sinnvoll, multidimensionale Indexe zu verwenden wie zum Beispiel **R-Tree-Indexe**. Diese haben den Vorteil, dass anders als bei klassischen Indexen, nur ein einziger Index durchsucht werden muss und nicht die Ergebnisse von allen einzelnen Dimensionen zusammengeführt werden müssen.