Pre-Caching hochdimensionaler Aggregate mit relationaler Technologie

Jens Albrecht, Marc Fiedler, Jürgen Görlich, Matthias Lemm und Thomas Ruf
{jens.albrecht, marc.fiedler, juergen.goerlich, matthias.lemm, thomas.ruf}@gfk.com
GfK Retail and Technology GmbH, Nordwestring 101, 90319 Nürnberg

Abstract: Die GfK Retail and Technology produziert Berichte für einige hundert Warengruppen auf Basis eines zentralen Data Warehouse. Die umfassenden und detailreichen, d.h. hochdimensionalen Auswertungen schränken die Einsatzfähigkeit von materialisierten Sichten stark ein. Cache-Techniken können aufgrund kontinuierlicher Änderungen am Datenbestand ebenfalls nicht eingesetzt werden. Um Zugriffe auf die Rohdaten dennoch vermeiden zu können, wurde ein kombinierter Ansatz entwickelt. Die benötigten Aggregate werden a priori in einen Aggregat-Cache geladen. Um eine hohe Wiederverwendbarkeit und einfache Wartbarkeit zu ermöglichen, werden die Aggregate nach den Hauptdimensionen zerlegt abgelegt. Das Verfahren unterstützt auch die Berechnung nicht-additiver Kennzahlen. Es wurde basierend auf einem relationalen Datenbanksystem realisiert und ist produktiv im Einsatz.

1 Einleitung

Die GfK-Gruppe ist eines der weltweit führenden Marktforschungsunternehmen mit einem Gesamtumsatz von über 1,1 Milliarden €. Im Geschäftsfeld „Retail and Technology“ stellt die GfK ihren Kunden umfassende Marktberichte zu technischen Gebrauchsgütern auf internationaler Basis zur Verfügung. Der Kundenkreis der GfK Retail and Technology (GfK RT) umfasst vornehmlich international operierende Markenartikelhersteller und Handelshäuser.

Die Basis des Reportings bilden Verkaufs-, Preis- und Bestandsdaten aus etwa 350.000 Geschäften in über 70 Ländern. Diese Daten werden im Rahmen eines umfangreichen ETL-Prozesses konsolidiert. Am Ende des Datenproduktionsprozesses werden die berichtsfertig konsolidierten Daten in die Auswertungsdatenbank, die Reporting Base, geladen. Die Reporting Base ist ein ROLAP-basiertes Data Warehouse auf Basis eines Oracle Real Application Clusters. Ein detaillierterer Überblick über das Auswertungssystem der GfK RT wird in [BG09] gegeben.

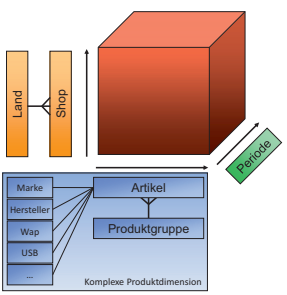


Abbildung 1: GfK RT Datenmodell

Der auswertungsbezogene Datenbestand ist nach einem Star-Schema organisiert. Das Datenmodell basiert auf 3 Hauptdimensionen, die Informationen zu Geschäften, Produkten und der Zeit enthalten (Ab-

bildung 1). Auf Basis dieser Dimensionen werden Fakten wie Verkaufsmenge, Preis und Bestand erhoben, aus denen verschiedene additive und nicht additive Kennzahlen (z.B. Umsätze, Marktanteile, Distributionen) berechnet werden können. Für die Auswertungen werden Geschäfte hierarchisch Ländern und Vertriebskanälen zugeordnet. Produkte sind nach Warengruppen, Kategorien und Sektoren organisiert. Eine Besonderheit ist die Vielzahl der dimensionalen Detailattribute, die insbesondere bei den Produkten für die Auswertung herangezogen werden. Abhängig von der Warengruppe werden bis zu 100 Warengruppen-spezifische Produktmerkmale erhoben. Neben der wichtigen Information über Marke/Hersteller können das bei Mobiltelefonen Eigenschaften wie WAP-Funktion, USB-Funktion und Kameraauflösung sein, bei Waschmaschinen hingegen Energieeffizienzklasse und Schleuderdrehzahl. Damit ergibt sich, dass der Auswertungsraum pro Warengruppe allein durch die Produktmerkmale potenziell 100-dimensional ist.

Aufgrund der Vorverdichtung auf Wochen- bzw. Monatsebene im Dateneingang ist die Größe des Data Warehouse mit knapp einem Terabyte noch überschaubar. Was die Datenverarbeitung in der GfK RT jedoch extrem anspruchsvoll macht, sind Masse und Komplexität der Berichte. Mehrere hundert Kunden erhalten nationale und internationale Standardberichte zu über 400 Warengruppen. Dafür werden jeden Monat über 100.000 Berichtsdateien und individualisierte Datenbanken produziert. Zusätzlich erhalten die Kunden einen Online-Zugang, der es ihnen ermöglicht, internationale Berichte online aus dem Data Warehouse System abzurufen.

Die Standardberichte zeichnen sich durch einen sehr hohen Detaillierungsgrad in Bezug auf die Produktmerkmale aus. Abbildung 2 zeigt einen typischen Seitenriss aus dem Bereich Mobiltelefone. Zu einem bestimmten marktrelevanten Aspekt der Warengruppe – in diesem Fall die eingebaute Kamera – sind hier die wesentlichen Merkmalsausprägungen dargestellt. Da verschiedene Merkmale in unterschiedlichen Kombinationen auftreten, sind in der Regel mehrere SQL-Anweisungen nötig, um einen Seitenriss zu berechnen.

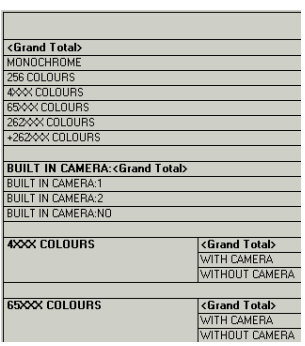


Abbildung 2: Beispiel eines Seitenrisses

Ein Standardbericht enthält zum Teil über 100 solcher Seitenrisse, die wiederum für bis zu 70 Länder pro Vertriebskanal ausgerechnet und meist zusätzlich noch nach Preisklassen unterteilt werden. Aufgrund der Kombinatorik kann die konventionelle Produktion eines Standardberichtes leicht über 100.000 SQL-Anweisungen umfassen, womit alle etablierten Systeme performancemäßig an ihre Grenzen stoßen. Online-Berichte werden im navigierenden Zugriff eher seitenweise abgerufen, so dass die Zahl der Anweisungen damit nicht so hoch ist. Die Antwortzeit-Anforderungen liegen dafür generell im Sekundenbereich, unabhängig vom Datenumfang. Sowohl die Batch-Produktion tausender Berichte als auch der Online-Zugang stellen somit enorme Herausforderungen in Bezug auf die Performance dar.

2 Performance-Optimierung durch redundante Datenstrukturen

In den vergangenen 10 Jahren wurde sehr viel zum Thema Performance-Optimierung in Data Warehouse Systemen veröffentlicht. Beschränkt man sich auf Methoden, die Performance durch Redundanz erzielen sollen, gibt es die Möglichkeit, entweder durch Vorberechnung von Aggregaten in Form materialisierter Sichten oder durch Cache-Techniken den Zugriff auf die Rohdaten zu vermeiden.

Materialisierte Sichten mit Anfragereformulierung haben den großen Vorteil, dass sie bereits in kommerziellen Datenbanksystemen wie Oracle verfügbar sind [AF06]. Eine besondere Stärke ist die universelle Wiederverwendbarkeit materialisierter Sichten. Die Auswahl geeigneter Aggregationsniveaus, aus denen viele Anfragen ableitbar sind, stellt jedoch ein Kernproblem dieses Ansatzes dar ([GL01, ACN00]). Aufgrund der Vielzahl der Auswertungsdimensionen in der GfK RT kommen pro Warengruppe bis zu 2^{100} verschiedene Aggregationsebenen in Betracht. Das würde entweder zu wenigen sehr feingranularen Materialisierungen oder einer großen Anzahl von Sichten gröberer Granularität zur Darstellung der häufigsten Merkmalskombinationen führen. De facto bedeutet das, dass in dieser Dimension nicht sinnvoll aggregiert werden kann. Nur die Geschäfte können auf Land- und Kanalebene voraggregiert werden, da deutlich weniger mit Geschäftsmerkmalen gearbeitet wird. Diese Art von Aggregationstabellen kann aber nur für die additiv berechenbaren Kennzahlen (Verkaufsmengen, Umsätze, Preise) eingesetzt werden. Die für die Marktforschung besonders wichtigen Distributionskennzahlen hingegen messen den Anteil der Geschäfte, die ein bestimmtes Produktsegment führen (z.B. Nokia-Telefone mit 3-Mega-Pixel-Kamera), so dass in keiner Dimension ein sinnvolles Aggregationsniveau für eine materialisierte Sicht gefunden werden kann. Da durch den Produktionsprozess in der GfK ständig neue Datenpakete pro Warengruppe, Land und Periode in die Reporting Base übertragen werden, würde die Verwendung von materialisierten Sichten oberhalb dieser Hierarchiestufen auch zu einem massiven Wartungsproblem führen.

Data Caching kann eingesetzt werden, um nahezu alle Arten von Anfragen zu beschleunigen. Allerdings müssen die nötigen Informationen erst einmal im Cache sein, und dann muss entscheidbar sein, unter welchen Voraussetzungen Anfragen auf Basis von Data Caches berechnet werden können. Typischerweise wird der Cache-Inhalt ähnlich einer materialisierten Sicht deskriptiv mit Hilfe einer definierenden Anfrage beschrieben [LGZ04]. Größter Vorteil des Cachings gegenüber materialisierten Sichten ist die dynamische Anpassung auf die Lastsituation [ABK+03]. Besonders problematisch ist hingegen die Entscheidung, wann der Cache invalidiert werden muss. Je höher das Aggregationsniveau des Anfrageergebnisses, desto häufiger muss es verworfen und wieder komplett neu berechnet werden. Der Result Cache in Oracle 11g verwirft den Cache-Inhalt beispielsweise bei jeder Änderung an einer der zugrundeliegenden Relationen, unabhängig davon, ob der Cache-Inhalt überhaupt betroffen ist.

Obwohl beide Techniken in verschiedensten Ausprägungen bereits Einzug in kommerzielle Datenbanksysteme gefunden haben, waren die damit erreichbaren Ergebnisse nicht ausreichend, um den Performance-Anforderungen des umfassenden und detailreichen Reportings in der GfK RT gerecht zu werden. Daraus ergab sich die Notwendigkeit, einen eigenen Ansatz zu finden, der die Stärken von vorberechneten Aggregaten mit Cache-

Techniken verbindet, ohne jedoch die Schwächen zu übernehmen. Erste Überlegungen in dieser Richtung wurden in [TAL06] veröffentlicht. Inzwischen wurde das Verfahren so weit entwickelt, dass es produktiv eingesetzt wird. Ziel der nachfolgenden Abschnitte ist es, einen Überblick über die Realisierung des Pre-Caching-Verfahrens und die damit erzielten Ergebnisse im praktischen Einsatz zu geben.

3 Pre-Caching anwendungsseitig vorberechneter Aggregate

In der GfK RT wurde mit der Aggregate Base ein Pre-Caching-System für die Realisierung folgender Design-Ziele entwickelt:

1. Die Aggregate müssen a priori bereit gestellt werden können. Ein rein dynamischer Ansatz hätte zur Folge, dass der erste Zugriff immer zu lange dauert. Da die Standardberichte periodisch mit den gleichen Segmentationen produziert werden und vorab bekannt sind, kann durch eine Vorab-Bereitstellung dieses Problem vermieden werden.
2. Es muss möglich sein, Aggregate dynamisch hinzuzufügen. Die Berichte ändern sich von Zeit zu Zeit, und es werden ad hoc neue Berichte erstellt. In diesem Fall muss das Verfahren fehlende Aggregate erkennen, nachberechnen und automatisch in den Cache aufnehmen.
3. Der Cache muss mit häufigen Änderungen umgehen können. In der GfK RT stellt die Kombination aus Land und Warengruppe eine Produktionseinheit dar, auf welcher pro Periode Daten in die Reporting Base eingespielt werden. Ein Aggregat, das vielleicht 50 Länder und 20 Warengruppen umfasst, würde damit in einer Berichtsperiode sehr häufig invalidiert.
4. Die Aggregate sollen in relationalen Strukturen abgelegt werden können, damit auf konventionelle Datenbanktechnologie mit ihren Vorteilen in Bezug auf Skalierbarkeit und Verfügbarkeit zurückgegriffen werden kann.

3.1 Das Aggregationsmodell

Um die Grundidee des Cache-Aufbaus und Zugriffs zu illustrieren, ist in Abbildung 3 oben ein einfacher Bericht dargestellt. Ausgegeben wird die Verkaufsmenge (*Sales Units*) der beiden Warengruppen *Mobilephones* und *Smartphones* in den Ländern Frankreich, England und Russland in der Periode März 2007. Neben der Gesamtsumme sind zusätzlich die Verkäufe nach Marken aufgeschlüsselt.

Um die Wiederverwendbarkeit und Wartbarkeit zu gewährleisten, werden die Aggregate nach den Hauptdimensionen Land, Produktgruppe und Periode zerlegt. Wie in Abbildung 3 unten angedeutet, wird beispielsweise der Gesamtwert für Nokia intern pro Land und Warengruppe berechnet und abgespeichert. Der Gesamtwert für den Bericht muss dann

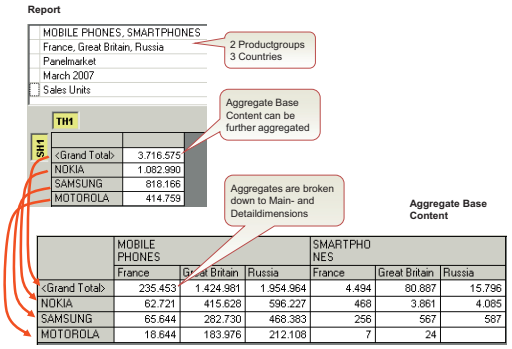


Abbildung 3: Berechnung von Kennzahlen mit Hilfe der Aggregate Base

nicht mehr aus Rohdaten, sondern nur noch aus 6 Aggregatzellen berechnet werden. Da in den Berichten beliebige Kombinationen von Produktmerkmalen gebildet werden können, müssen diese gesondert behandelt werden. Jede Aggregatzelle ist neben den Hauptdimensionen beschrieben durch ein Tupel von (Attribut:Wert)-Pärchen, z.B. (Marke:Nokia) oder (Marke:Nokia;UMTS:Ja;Kamera:Ja). Diese potenziell hochdimensionalen Merkmalskombinationen werden auf einen eindimensionalen Schlüssel abgebildet. Damit können die heterogen strukturierten Aggregate in einer einzigen relationalen Faktentabelle abgelegt werden.

Formal ausgedrückt wird in der Aggregate Base ein standardisierter Satz von Basiskennzahlen $F = (F_1, \ldots, F_f)$ auf Granularität von Haupt- und Detaildimensionsattributen abgelegt. Die Hauptdimensionen der Aggregate Base M werden gebildet aus geeigneten Kategorienattributen eines relationalen Star- Schemas: $M = \{M_1, \ldots, M_m\}$. In der GfK ist $m = 3$ und die Kategorienattribute sind Land, Produktgruppe und Periode.

Die Detaildimensionen D dienen der Darstellung semi-strukturierter Detailinformationen von beliebigen Kombinationen der Merkmale aus den Hauptdimensionen: $D = \{D_1, \ldots, D_d\}$ mit $d \leq m$. Über eine Abbildungsrelation k wird jede mögliche Kombination von (Attribut:Wert)-Pärchen der Detaildimensionen auf genau einen numerischen Schlüssel abgebildet. k ist also eine bijektive Funktion:

$$k\ \{(\text{Attribut:Wert})_1 \ldots (\text{Attribut:Wert})_n\} \rightarrow \mathbb{N}$$

Für die GfK gilt $d = 2$, da nur Merkmalskombinationen aus der Produkt- und Geschäftsdimension berücksichtigt werden (PKey, SKey). Für eine Datenzelle aus dem Beispiel (Abbildung 3) ergibt sich der folgende Eintrag in die Aggregate Base:

$$\underbrace{(\text{France}}_{\text{Land}}, \underbrace{\text{Mobilephones}}_{\text{Produktgruppe}}, \underbrace{\text{March 2007}}_{\text{Periode}}, \underbrace{337439}_{\text{PKey: k((Marke:Nokia))}}, \underbrace{337434}_{\text{SKey: k((Grand Total))}}, \underbrace{62721}_{\text{Sales Units}}, \ldots)$$

Das Aggregationsmodell der Aggregate Base reduziert somit die Dimensionalität eines $m + r$-dimensionalen Datenraumes auf $m + d$ Dimensionen mit $d << r$. [1] Durch diese

[1] m Anzahl der Hauptdimensionen, r Anzahl der Merkmalsattribute, d Anzahl der Detaildimensionen

Art von Dimensionsreduktion wird die Modellierung durch ein kompaktes relationales Datenschema der Form (M, D, F) ermöglicht.

3.2 Berechnung von Kennzahlen auf Basis der Aggregate Base

Erster Schritt beim Zugriff auf Kennzahlen der Aggregate Base ist die Übersetzung des Auswertungsraumes eines Berichtes auf die Hauptdimensionswerte des Datenraumes der Aggregate Base (Addressierung). Im Beispiel aus Abbildung 3 werden zunächst die Warengruppen (Mobilephones und Smartphones), die Länder (Frankreich, England und Russland) und die Periode (März 2007) aufgelöst. Im zweiten Schritt müssen die Detailinformationen, z.B. (Marke:Nokia) aus der Produktinformation und der *Grand Total* aus der Geschäftsdimension auf ihre korrespondierenden numerischen Werte übersetzt werden (*Key Mapping*).

Aggregate können sowohl direkt abgegriffen werden (Punktzugriff), z.B. die Verkäufe für Mobiltelefone der Marke Nokia in Frankreich im März 2007, als auch weiter verdichtet werden, wie beispielsweise das Länder- und Warengruppenkumulat für die Marke Nokia (Abbildung 3). Die weitere Verdichtung von hinterlegten Aggregaten findet unter Nutzung dimensionaler Hierarchien oder über frei definierbare Gruppen für die Attribute der Hauptdimensionen statt. Eine Aggregation über die Attribute der Detaildimensionen ist nicht möglich, da der Schlüsselwert k auch die Werteausprägungen kapselt. Jede weitere Verdichtung der materialisierten Basiskennzahlen F wird ausschließlich über die Aggregationsfunktion SUM realisiert, auch wenn die eigentliche Urberechnung der Basiskennzahlen nicht notwendigerweise additiv ist. Eine nicht-additive Metrik ist beispielsweise die *Distribution*. Die Distribution prozentuiert die Anzahl von Geschäften, die Produkte mit bestimmten Merkmalen führen, auf die Anzahl aller beobachteten Geschäfte. Diese beiden Basiskennzahlen werden in der Aggregate Base abgelegt. Auch wenn die Basiskennzahlen nicht additiv berechenbar sind, so können die Aggregatzellen der Aggregate Base zumindest über Länder hinweg summiert werden, da Geschäfte eindeutig einem Land zugeordnet werden und somit deren Anzahl überschneidungsfrei weiter verdichtet werden kann.

Mit Hilfe der Aggregate Base können Reports mit beliebig heterogenen Seitenrissen (unterschiedlichste Kombinationen von Produkt- und Geschäftsmerkmalen) auf Basis sehr weniger SQL-Anweisungen komplett berechnet werden. Im einfachsten Fall (Berechnung der Verkaufsmenge) ist bei Zugriff auf die Aggregate Base lediglich eine SQL-Anweisung nötig, die Aggregate über die Hauptdimensionsattribute verdichtet und auf die PKey- und SKey-Werte, die durch den Seitenriss addressiert werden, einschränkt. Im Beispiel aus Abbildung 3 werden insgesamt 4 PKeys und 1 SKey (*Grand Total*) für die angefragten Hauptdimensionsattribute aus der Aggregate Base angefordert. Der gesamte Report wird auf Basis von

$$\underbrace{3}_{\text{Land}} \times \underbrace{2}_{\text{Produktgruppe}} \times \underbrace{1}_{\text{Periode}} \times \underbrace{4}_{\text{PKey}} \times \underbrace{1}_{\text{SKey}} = 24$$

Aggregaten der Aggregate Base berechnet.

Beim Zugriff auf die Aggregate Base wird ein optimistischer Ansatz mit „Vollständigkeitsvorbehalt“ realisiert, d.h. bei jedem Zugriff auf Aggregate wird davon ausgegangen, dass diese bereits vorhanden sind. Eine Kennzahl kann nur dann im Reporting weiter verwendet werden, wenn sie das Kriterium der Vollständigkeit erfüllt. Ein Aggregat ist genau dann vollständig aus der Aggregate Base berechenbar, wenn die Ergebniskardinalität gleich der Erwartungskardinalität ist. Für die Vollständigkeitsprüfung müssen also Ergebniskardinalität und Erwartungskardinalität ermittelt werden. Die Ergebniskardinalität wird über einen einfachen COUNT(*) direkt beim Zugriff auf die Aggregate Base ermittelt. Für die Bestimmung der Erwartungskardinalität wird eine Kontroll-Relation herangezogen. Diese ist eine *Factless Fact Table*, die nur aus den Hauptdimensionen $M = \{M_1, \ldots, M_m\}$ der Aggregate Base besteht. In dieser Relation wird hinterlegt, ob für eine Kombination aus (Land, Produktgruppe, Periode) Rohdaten geladen sind. Ist das der Fall, so muss auch ein entsprechendes Aggregat vorhanden sein. Die Kontroll-Relation ist extrem kompakt, so dass die Zugriffskosten für die Bestimmung der Erwartungskardinalität völlig vernachlässigbar sind.

In dem Bericht aus Abbildung 3 ist der Erwartungswert für jede Ergebniszelle 6 (3 Länder $\times$ 2 Warengruppen $\times$ 1 Periode). Würden die Aggregatzellen von Mobilephones/Frankreich aus irgendeinem Grund fehlen, wäre die Ergebniskardinalität 5 und der mit der Aggregate Base errechnete Wert unvollständig. In einem solchen Fall werden die entsprechenden Aggregate dynamisch nachgerechnet und in die Aggregate Base eingetragen.

3.3 Befüllung der Aggregate Base

Die Vorberechnung von Aggregaten für die Aggregate Base ist Bestandteil des Ladeprozesses für die Rohdaten. Dazu wird ein speziell definierter Versorgungsbericht verwendet, der die im Standard-Reporting verwendeten Seitenrisse enthält. Da die Daten periodisch auf Ebene von Land und Produktgruppe geladen werden, findet das Berechnen der Aggregate auf einem sehr überschaubaren Datenbestand statt. Tests haben erwiesen, dass bereits bei einmaliger Berechnung eines komplexen Berichtes das Pre-Caching lohnenswert ist. Die Zeit für die Berechnung der Aggregate und für die Auswertung des Berichtes ist insgesamt geringer, als wenn der Bericht vollständig auf Basis der Rohdaten berechnet wird. Die initiale Vorberechnung amortisiert sich insbesondere dann, wenn Aggregate sehr häufig wiederverwendet werden können.

3.4 NULL-Wert-Kompression bei Verwendung mehrerer Detaildimensionen

Die Berechnung der Erwartungskardinalität für die Vollständigkeitsprüfung auf Grundlage der Hauptdimensionsattribute erfordert zusätzlich die Speicherung von NULL-Aggregaten. NULL-Aggregate entstehen bei der Befüllung der Aggregate Base immer dann, wenn für eine bestimmte Merkmalskombination keine Rohdatensätze existieren. Das trifft beispielsweise für Marken zu, die in einem Land nicht verkauft werden. In der Aggregate Base

werden diese NULL-Werte permanent abgelegt. Somit ist sichergestellt, dass die Ergebniskardinalität bei Vollständigkeit auch tatsächlich korrekt ermittelt werden kann.

Das direkte Ablegen von NULL-Aggregaten ist bei einer Detaildimension noch akzeptabel, im mehrdimensionalen Fall führt dieses Vorgehen jedoch zu einem nicht tolerierbaren Anwachsen des Datenvolumens. Bei der GfK RT ist die Detaildimension zweidimensional und besteht aus PKey und SKey. Typische Bestandteile des SKeys sind Geschäftsmerkmale wie der Vertriebskanal oder die Umsatzgrößenklasse. Da in den Standardberichten alle Produktmerkmalskombinationen mit allen Geschäftsmerkmalskombinationen ausmultipliziert werden, ist der errechnete Datenraum beträchtlich. Untersuchungen haben gezeigt, dass über 90% der Aggregatzellen NULL-Werte enthalten.

Um nicht für jede mögliche Kombination von PKey- und SKey-Werten ein NULL-Aggregat speichern zu müssen, werden sämtliche Werte der Detaildimensionen zur NULL-Wert-Speicherung linearisiert (Abbildung 4). Statt des kompletten Datenraumes werden also in einer gesonderten Relation nur zwei Listen für die Detaildimensionen abgelegt, aus denen sich der leere Datenraum wieder aufspannen lässt. Dieser Raum leerer Aggregatzellen wird dann mit den Aggregatzellen, die tatsächlich Zahlenwerte beinhalten, überlagert und zusammenaggregiert. Die Rekonstruktion ist zwar nicht kostenlos, aber deutlich günstiger als der Zugriff auf unkomprimierte NULL-Aggregate.

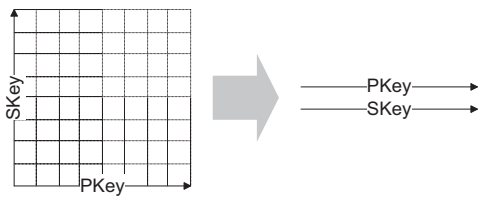
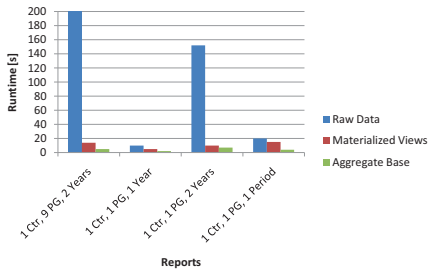


Abbildung 4: Zweidimensionale NULL-Wert-Kompression

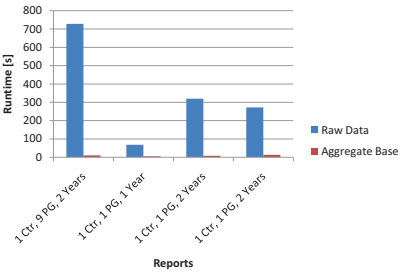
Die Kompressionsrate ist umso größer, je mehr Aggregatzellen auf einmal ausgerechnet werden. Werden 1000 PKey-Werte mit 50 SKey-Werten kombiniert, so werden für die Information, dass diese 1000×50 Zellen berechnet wurden, lediglich 1050 NULL-Werte abgelegt. Werden nur 2 PKey-Werte und 2 SKey-Werte kombiniert, sind für diese 4 Zellen auch in der linearisierten Darstellung 4 NULL-Werte nötig. Zusätzlich müssen noch die Zellen mit Zahlenwerten abgelegt werden. In diesem Fall ist die Linearisierung also sogar schädlich. Die NULL-Wert-Komprimierung ist deswegen besonders profitabel bei der initialen Befüllung der Aggregate Base durch einen umfangreichen Versorgungsbericht. In diesem Fall werden leicht Kompressionsraten von über 90% erreicht. Für dynamische Nachberechnungen wird auf die NULL-Wert-Kompression hingegen verzichtet.

3.5 Evaluation

Ein wesentliches Ziel bei der Entwicklung der Aggregate Base war es, eine Performancesteigerung für das interaktive Reporting zu erreichen. Abbildung 5 a) stellt die Laufzeiten unterschiedlich komplexer Berichte bei der Berechnung von Verkaufswerten gegenüber. Für den Vergleich mit materialisierten Sichten wurde eine Aggregationstabelle angelegt, welche die Geschäftsdimension auf Länder-Ebene beschränkt auf wenige häu-



a) Laufzeit für Sales Units



b) Laufzeit für Distribution

Abbildung 5: Laufzeiten unterschiedlicher Berichte auf Basis von Rohdaten, materialisierter Sichten und Aggregate Base

fige Geschäftsmerkmale hoch aggregiert vorhält, auf der Produktebene aber auf Basisgranularität bleibt, um alle Merkmalskombinationen in den Berichten abdecken zu können. Während die Laufzeiten für die Urberechnung von Kennzahlen aus Rohdaten für den interaktiven Betrieb nicht akzeptabel sind, kann bei Verwendung von vorverdichteten Daten, insbesondere bei sehr umfangreichen Fusionen aus Ländern und/oder Warengruppen und längeren Zeiträumen, eine erhebliche Performancesteigerung erzielt werden. Die Berechnung aus der Aggregate Base ist meist um Faktor 2-3 schneller als die Berechnung basierend auf materialisierten Sichten. Dieser Performancevorteil resultiert aus der Tatsache, dass im Gegensatz zu materialisierten Sichten auch für einen komplexen heterogen strukturierten Bericht meist nur eine einzige SQL-Anweisung für die Berechnung aus der Aggregate Base notwendig ist. Hinzu kommt, dass nur eine begrenzte Anzahl von Anfragen tatsächlich durch materialisierte Sichten unterstützt werden kann. Ursache dafür ist das starre relationale Schema materialisierter Sichten, welches deren universelle Nutzbarkeit einschränkt. Je komplexer die Urberechnung von Kennzahlen ist, desto höher ist der Nutzen der Aggregate Base. Die Berechnung von Distributionen ist ein solches Beispiel (Abbildung 5 b)). Die Distributionsberechnung ist zum einen besonders aufwändig und zum anderen nicht durch Nutzung materialisierter Sichten optimierbar. Im Vergleich zur Urberechnung ist bei gefüllter Aggregate Base eine Beschleunigung um das bis zu 100-fache erreichbar.

4 Zusammenfassung

Für die speziellen Anforderungen in der GfK Retail and Technology wurde ein neuartiges Pre-Caching-Verfahren entwickelt. Vorab bekannte Aggregate für die Produktion von Standardberichten werden mit der Bereitstellung der Rohdaten automatisch berechnet und in die Aggregate Base eingefügt, damit bereits der erste Zugriff schnell erfolgen kann. Fehlende Aggregate werden zur Berichtslaufzeit durch eine automatische Vollständig-

keitsprüfung erkannt und können dynamisch ergänzt werden.

Die Aggregate werden in relationalen Strukturen in einer Oracle-Datenbank abgelegt. Die Verwendung relationaler Strukturen für die Speicherung von Aggregaten, die für beliebige Merkmalskombinationen aus der Produkt- und Geschäftsdimension berechnet werden, setzt eine Dimensionsreduktion des Datenraumes voraus. Dadurch wird die Speicherung und der Zugriff auf Aggregate beliebiger Granularität in den Detaildimensionen ermöglicht. Der Zugriff erfolgt anwendungsseitig über einen in Eigenentwicklung entstandenen Analyse-Server, der sowohl vom Online-Frontend angesprochen werden kann als auch die Batch-Produktion von Berichten ermöglicht.

Das System ist produktiv im Einsatz. Die erreichten Laufzeitverbesserungen reichen je nach Berichtsumfang und Komplexität von Faktor 2 bis zu einigen Größenordnungen.

Literatur

[ABK+03] Mehmet Altinel, Christof Bornhövd, Sailesh Krishnamurthy, C. Mohan, Hamid Pirahesh und Berthold Reinwald. Cache Tables: Paving the Way for an Adaptive Database Cache. In *Proceedings of 29th International Conference on Very Large Data Bases, (VLDB 2003, September 9-12, Berlin, Germany)*, Seiten 718–729, 2003.

[ACN00] Sanjay Agrawal, Surajit Chaudhuri und Vivek R. Narasayya. Automated Selection of Materialized Views and Indexes in SQL Databases. In *Proceedings of 26th International Conference on Very Large Data Bases, (VLDB 2000, September 10-14, Cairo, Egypt)*, Seiten 496–505, 2000.

[AF06] Jens Albrecht und Marc Fiedler. Datenbank-Tuning – einige Aspekte am Beispiel von Oracle 10g. *Datenbank-Spektrum*, 16:26–33, 2006.

[BG09] Andreas Bauer und Holger Günzel, Hrsg. *Data-Warehouse-Systeme Architektur, Entwicklung, Anwendung*. dpunkt.verlag, 3. Auflage, 2009.

[GL01] Jonathan Goldstein und Per-Åke Larson. Optimizing Queries Using Materialized Views: A Practical, Scalable Solution. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, (SIGMOD 2001, May 21-24, Santa Barbara, USA)*, Seiten 331 – 342, 2001.

[LGZ04] Per-Åke Larson, Jonathan Goldstein und Jingren Zhou. MTCache: Transparent Mid-Tier Database Caching in SQL Server. In *Proceedings of the 20th International Conference on Data Engineering, (ICDE 2004, March 30 - April 2, Boston, USA)*, Seiten 177–189, 2004.

[TAL06] Maik Thiele, Jens Albrecht und Wolfgang Lehner. Optimistic Coarse-Grained Cache Semantics for Data Marts. In *Proceedings of the 18th International Conference on Scientific and Statistical Database Management, (SSDBM 2006, July 3-5, Vienna, Austria)*, Seiten 311–320, 2006.