Wie viele Dimensionen hat ein Würfel?

Der Einsatz von Datamarts als Analysehilfen in einem Data Warehouse

Andrea Kennel

Beim Start eines Data Warehouse-Projekts gibt es oft das Problem, dass Informatiker und Benutzer nicht dieselbe Sprache sprechen. Der Benutzer weiss nicht genau, was er sich vorstellen soll und der Informatiker weiss nicht, was der Benutzer genau braucht. Für einen guten Projektstart ist es daher wichtig, dass alle Beteiligten wissen, wovon sie sprechen. Dieser Artikel ist genau für diesen Einstieg gedacht. Die Grundbegriffe werden anschaulich und verständlich mit Beispielen erklärt. Das Hauptgewicht liegt dabei bei den Auswertungsmöglichkeiten, da Auswertungen das Hauptziel eines Data Warehouses sein sollten.

1 Ist ein Datamart ein Data Warehouse?

1.1 Data Warehouse

Data Warehouse heisst auf deutsch Daten-Lagerhaus. Der Vergleich mit einem Lagerhaus ist recht treffend. In einem normalen Lagerhaus werden in grossen Mengen Waren gelagert, auf die möglichst einfach wieder zugegriffen werden muss. In einem Daten-Lagerhaus oder Data Warehouse werden grosse Mengen von Daten gelagert, die später wieder gelesen werden sollen. Das Ziel eines Data Warehouses ist nicht primär das Lagern von Daten, sondern das Sammeln der Daten, um sie nach verschiedenen Kriterien auswerten zu können. Im Vergleich zu einem operativen System werden in einem Data Warehouse normalerweise Daten über längere Zeit (mehrere Jahre) gesammelt, um Auswertungen über die Zeit vornehmen zu können.

1.2 Datamart

Ein Datamart kann mit Daten-Marktstand übersetzt werden. Diese Metapher eines Markstandes ist treffend. In einem Datamart will man nicht alle Daten sehen, sondern nur eine spezielle Auswahl. Ein Datamart ist somit immer nur eine Auswahl von Daten, die für einen bestimmten Benutzerkreis gedacht ist. So wie es Marktstände für Gemüse und Käse gibt, gibt es Datamarts, die verdichtete Zahlen für das Management bereitstellen oder Datamarts, die Bereichszahlen für Bereichsleiter zur Verfügung stellen.

Durch eine gezielte Analyse der Geschäftsprozesse lassen sich Kerninformationen herauskristallisieren, so dass Datamarts aufgebaut werden können, die mit 20% der gesamten Daten eines Warehouses 80% der Anfragen abdecken.

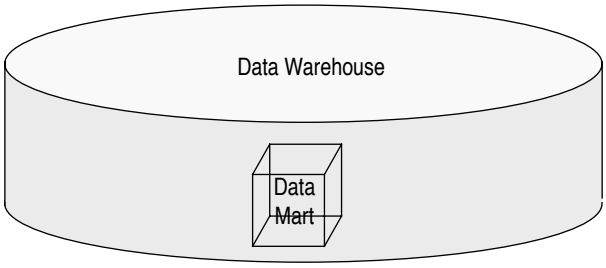

Abb. 1: Ein Datamart ist ein bestimmter Ausschnitt aus den Daten eines Data Warehouses

1.3 Vergleich

Ein Data Warehouse ist eine Sammlung von Daten über die Zeit. Diese Sammlung ermöglicht verschiedene Auswertungen.

Ein Datamart ist ein bestimmter Ausschnitt aus den Daten eines Data Warehouses (Abb. 1).

2 Die Sicht des Benutzers auf das Data Warehouse

2.1 Relationale Sicht

Die Daten in einem Data Warehouse sind relational in Form von Datenbanktabellen abgelegt. Nehmen wir als Beispiel eine Warenhauskette, die unter anderem Campingartikel verkauft. Im Data Warehouse ist abgelegt, wann wo wie viele Campingartikel verkauft wurden. Dazu werden auch noch weitere Daten

Andrea Kennel, Dr. sc. techn., ist als Consultant bei der Datenbankberatungsfirma Trivadis AG tätig. Dabei ist sie in unterschiedlichen Bereichen tätig. Neben Unterstützung im Aufbau von Datamarts entwickelt sie auch Datenbankapplikationen und hilft bei der Migration von Daten. Wie einige Consultants der Firma Trivadis gibt Frau Kennel ihr Wissen auch in Kursen, die von der Trivadis angeboten werden, weiter, was ihr dank ihrer didaktischen Zusatzausbildung leicht fällt.

wie Einheiten und Kosten abgelegt. Die entsprechende Tabelle sieht folgendermassen aus:

Produkt	Region	Zeit	Verkauf	Kosten	Einheiten
...	...	...	...	...	...
Zelt	Bern	07 1998	9'200	8'720	30
Zelt	Zürich	06 1998	13'721	12'052	50
Zelt	Zürich	07 1998	15'574	14'925	60
...	...	...	...	...	...

2.2 Multidimensionale Sicht

Die obige relationale Darstellung ist gut für die Datenhaltung, nicht aber für den Benutzer. Die Liste, die der Benutzer sehen will und interpretieren kann, sieht etwas anders aus. Folgend zwei mögliche Beispiele:

Verkauf in der Region Zürich					
	06 1998	07 1998	...	...	...
Zelt	13'721	15'574	...	...	...
...	...	...	...	...	...

Region Zürich						
	Verkauf		Kosten		Einheiten	
	06 1998	07 1998	06 1998	07 1998	06 1998	07 1998
Zelt	13'721	15'574	12'052	14'925	50	60
...	...	...	...	...	...	...

Diese beiden Beispiele stellen eine multidimensionale Sicht dar. Im ersten Beispiel haben wir in der x-Achse die Zeit und in der y-Achse die Produkte. In der z-Achse ist nur eine Scheibe sichtbar. Darin haben wir die Region und Wert, wobei festgelegt ist, dass wir für die Region Zürich den Wert Verkauf sehen wollen. Im unteren Beispiel ist die Dimension Wert zusätzlich zur Dimension Zeit in die x-Achse geholt worden. Somit werden in einer Achse zwei Dimensionen dargestellt.

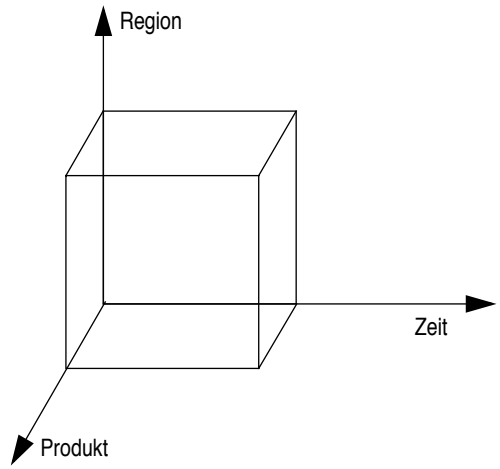

Abb. 2: Darstellung als Würfel

Diese Art der Darstellung mit Dimensionen wird auch multidimensionale Darstellung oder Würfel genannt. Betrachten wir einmal nur die Verkaufszahlen. Lassen wir die Dimension Zeit in der x-Achse, die Dimension Produkt in der y-Achse und die Dimension Region in der z-Achse. So spannen diese drei Dimensionen genau einen Würfel auf. Jede Zelle des Würfels enthält genau einen Wert, der den Verkaufswert für ein Produkt in einer Region zu einem Zeitpunkt angibt (Abb. 2 und Abb. 5).

2.3 Drill Down und Drill Up

Nun interessiert normalerweise nicht nur welche Verkäufe mit Zelten gemacht wurden. Sind einige Zahlen auffällig gut oder schlecht, so möchte man auch wissen, wie sich diese zusammensetzen. Beispielsweise möchte man wissen, ob der unerwartete Erfolg wegen der neuen Igluzelte erzielt wurde oder wegen der Standardzelte. Somit muss es möglich sein in das Produkt Zelt, das sich aus Igluzelten und Standardzelten zusammensetzt, genauer hinein zu sehen. Diese Funktionalität des Aufklappens wird Drill Down genannt. Ein Beispiel einer aufgeklappten Darstellung:

Verkauf in der Region Zürich					
	06 1998	07 1998	...	...	...
- Zelt	13'721	15'574	...	...	...
Iglu	6'520	8'215			
Standard	7'201	7'359			
...	...	...	...	...	...

Analog zum Aufklappen gibt es auch das Zuklappen oder Drill Up. Damit über jede Dimension bis auf ein Gesamttotal zugeklappt werden kann, muss für jede Dimension eine Hierarchie definiert sein, die festlegt, wie die Daten zu verdichten sind (Abb. 3).

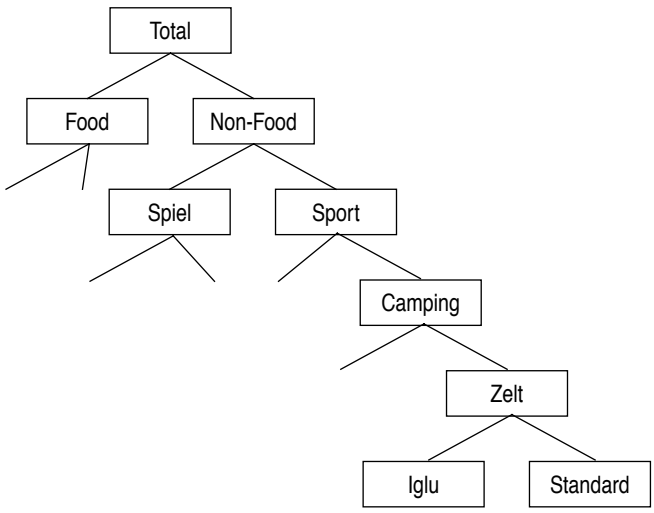

Abb. 3: Beispiel einer Hierarchie über der Dimension Produkt

2.4 Charakteristiken von OLAP

Die vier Buchstaben OLAP stehen für OnLine Analytical Processing. Das bedeutet, dass der Endbenutzer, der die Daten analysieren will, direkt auf die Daten zugreifen kann und ad-hoc-Auswertungen definieren kann. Damit der Endbenutzer auf möglichst einfache Art seine Abfragen formulieren kann, müssen die Daten und ihre Strukturen so dargestellt sein, dass sie einfach verständlich sind. Dies ist mit der multidimensionalen Sicht möglich. Der Benutzer legt fest, welche Dimensionen ihn interessieren, was er in der x-Achse und was in der y-Achse sehen möchte. Weiter kann der Benutzer in jeder Dimension festlegen, an welchen Dimensionswerten er interessiert ist. So kann er beispielsweise angeben, dass er die Dimensionen Zeit und Produkt ganz sehen will und die anderen Dimensionen auf Region Zürich und Wert Verkauf einschränken will. So kann die Abfrage über die Dimensionen festgelegt werden.

Für eine Analyse interessieren meistens nicht nur Detailzahlen, sondern vor allem Verdichtungen. Bei Auffälligkeiten in den verdichteten Daten oder für Stichproben ist das Hinuntergehen auf detailliertere Daten von Interesse. Dies ist ein wichtiger Bestandteil von OLAP und entspricht dem Drill Down und Drill Up.

Zusammengefasst ergeben sich folgende Charakteristiken für OLAP:

- Abfrage über Dimensionen
  - Dimensionen der Darstellung wählen
  - Dimensionswerte einschränken
- Drill Down, Drill Up
  - Vom Total zum Detail
  - Vom Detail zum Total

3 MOLAP

3.1 Funktionsweise von MOLAP

MOLAP steht für multidimensionales OLAP. Das bedeutet, dass die Daten nicht nur für den Benutzer multidimensional dargestellt werden, sondern auch so gespeichert werden. Dazu werden normalerweise die Daten aus dem relationalen Data Warehouse in eine multidimensionale Datenbank kopiert.

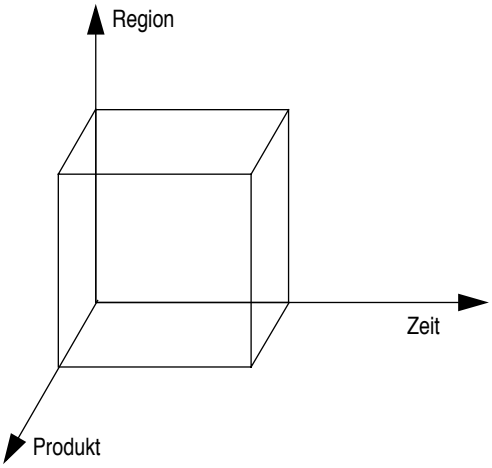

Abb. 4: Basiswürfel

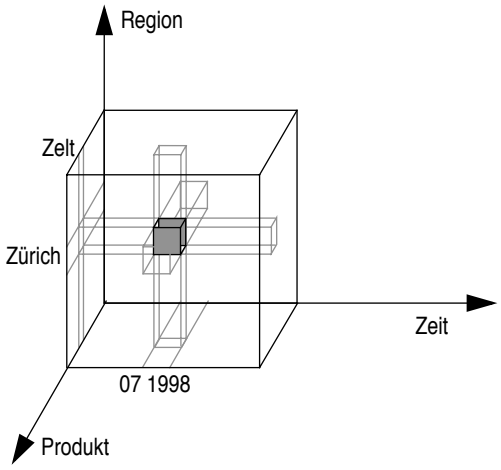

Abb. 5: Einzelne Zelle im Würfel

Die Funktionsweise von MOLAP kann am einfachsten anhand eines Beispiels mit den drei Dimensionen Produkt, Zeit und Region veranschaulicht werden. Die Werte, die in diesem Würfel gespeichert sind, sind Verkäufe. Im Basiswürfel sind die Basisdaten wann wo was verkauft wurde gespeichert (Abb. 4).

Jede einzelne Zelle des Würfels ist durch die Dimensionswerte bestimmt. So kann auf Grund der Dimensionswerte die entsprechende Zelle adressiert werden. So erhalten wir beispielsweise über die Dimensionswerte “Zelt”, “Zürich” und “07 1998” den entsprechenden Verkaufswert (Abb. 5).

Damit Verdichtungen berechnet werden können, müssen über den Dimensionen Hierarchien definiert sein. So können Monate zu Quartalen, Halbjahren und Jahren verdichtet werden. Über der Dimension Produkt kann eine Hierarchie definiert sein, die besagt, welche Produkte zu Zelten zusammengefasst werden und welche Produkte oder Produktegruppen zu Camping gehören.

Entlang dieser Hierarchien werden dann Verdichtungen berechnet und direkt in einem erweiterten Würfel abgespei-

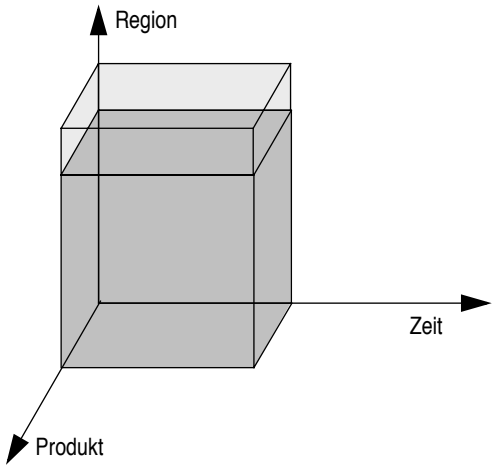

Abb. 6: Erweiterung des Basiswürfels in einer Dimension

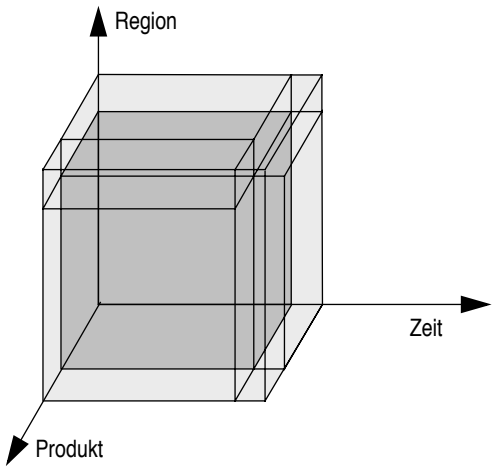

Abb. 7: Erweiterung des Basiswürfels in allen Dimensionen

chert. Eine Verdichtung über die Dimension Region ergibt somit eine Erweiterung des Würfels auf der Regionen-Achse (Abb. 6).

Nach dem Verdichten über alle Dimensionen erhalten wir einen grösseren Würfel (Abb. 7).

Die neuen Zellen des Würfels werden genau wie die Zellen des Basiswürfels über die Dimensionswerte angesprochen. Die Dimensionswerte sind dann einfach auf einer anderen Hierarchiestufe. So kann der Verkaufswert für das dritte Quartal 1998 für die Campingartikel in der Region Deutschschweiz abgefragt werden.

3.2 Möglichkeiten und Grenzen

Dadurch, dass mit MOLAP alle möglichen Verdichtungen über die vorgegebenen Hierarchien bereits vorgerechnet wurden, ist der Zugriff auf Daten recht schnell. Obwohl die Verdichtungen vorgegeben sind, ist es möglich, manuell einzelne Dimensionswerte zu gruppieren und ad-hoc Summen zu bilden. Da dabei normalerweise nur wenige Zahlen summiert werden müssen, ist die Antwortzeit auch dann akzeptabel.

Eine klare Grenze von MOLAP liegt aber bei der Zeit, die für die Vorberechnungen benötigt wird. Alle Verdichtungen werden im voraus berechnet und im Würfel abgelegt. Oft wird dies jede Nacht gemacht, da auch die Daten vom Vortag interessieren. Somit ist die Zeit für die Vorberechnungen beschränkt. Je nach Datenmenge und vor allem je nach Anzahl der Dimensionen kann eine Vorberechnung aber länger als 24 Stunden dauern.

Wird das Data Warehouse laufend mit aktuellen Daten gefüllt, so fliessen diese nicht laufend in den Würfel. Damit können im Würfel keine aktuellen Daten dargestellt werden, sondern nur Daten bis zum letzten Laden.

Somit ist der Einsatz von MOLAP dann sinnvoll, wenn:

- nur ein Ausschnitt der Daten zur Verfügung gestellt wird,
- nur eine überschaubare Anzahl von Dimensionen dargestellt und ausgewertet werden,
- keine aktuellen Daten benötigt werden.

Die Beschränkung auf 4 bis 6 Dimensionen ist meistens auch für den Endbenutzer von Vorteil, da die Komplexität sonst auch für ihn zu gross wird.

4 Flexiblere Möglichkeiten mit ROLAP

4.1 Funktionsweise von ROLAP

Um die Grenzen von MOLAP zu überwinden, gibt es ROLAP. Dabei steht das R für relational. Obwohl die Daten dem Endbenutzer multidimensional dargestellt werden, werden sie relational gespeichert. Da die Daten in einem Data Warehouse normalerweise relational gespeichert sind, kann ROLAP direkt auf dem Data Warehouse aufsetzen und muss keine Daten kopieren. Die Verdichtungen werden dann berechnet, wenn sie gebraucht werden. Die entsprechenden Werkzeuge generieren aus den Anfragen der Endbenutzer Datenbankabfragen in SQL.

4.2 Möglichkeiten und Grenzen

Da die Daten ad-hoc berechnet werden und nicht vorverdichtet werden, ist die Datenmenge und die Anzahl der Dimensionen weniger kritisch. Auch grosse Datenmengen können bewältigt werden, und vor allem können auch aktuelle Daten berücksichtigt werden. Der Nachteil liegt bei der Antwortzeit. Komplexere Abfragen bei denen viele Verdichtungen berechnet werden müssen, können etwas länger dauern.

4.3 Unterschiede MOLAP, ROLAP

Die Unterschiede zwischen MOLAP und ROLAP werden in der folgenden Tabelle zusammengefasst:

	MOLAP	ROLAP
Vorberechnung	lange	keine
Lesen verdichteter Daten	schnell	langsam
Datenmenge und Anzahl Dimensionen	eingeschränkt	flexibel

4.4 HOLAP

Seit einiger Zeit ist neben den Begriffen MOLAP und ROLAP nun auch der Begriff HOLAP aufgetaucht. Das H steht für hybrid, was bedeutet, dass HOLAP eine Mischung zwischen MOLAP und ROLAP ist. Konkret werden häufig gebrauchte Verdichtungen wie in MOLAP vorberechnet und gespeichert. Je nach Abfrage werden die Verdichtungen wie in ROLAP berechnet oder eben aus der vorgerechneten Tabelle gelesen.

5 Zusammenfassung

Meistens sind in einem Data Warehouse so viele Daten gespeichert, dass der Benutzer vor lauter Bäumen den Wald nicht mehr sieht, oder eben vor lauter Daten keine Information findet. Daher braucht es Werkzeuge, die den Benutzer in der Analyse unterstützen.

Je nach Bedürfnissen des Benutzers ist es wichtig, dass möglichst alle Daten möglichst flexibel analysiert werden können. Solche Werkzeuge können mit einem Buschmesser verglichen werden, mit denm der Benutzer sich durch den Datendschungel kämpfen kann und alles sehen kann, was er will. Dabei besteht aber die Gefahr, dass er nicht alles findet, was er sucht.

Oft ist es sinnvoll, wenn der Benutzer nicht alle Daten sieht, sondern nur die, die für ihn sicher relevant sind. Solche Werkzeuge führen den Benutzer mehr, so dass er sich wie auf einer Autobahn durch seine Daten bewegen kann. Dabei findet er schneller zum Ziel. Dafür besteht die Gefahr, dass er verschiedene Details, die nicht direkt an der Autobahn sind, nicht sieht.

Daher ist es wichtig, sich genau zu überlegen, welche Information aus einem Data Warehouse herausgelesen werden soll. Dann kann entschieden werden, welche Art von Datamart zum Einsatz kommen soll und wie viele Würfel mit wie vielen Dimensionen eingesetzt werden.

Call for Contribution

Themenheft 6/1999
Software-Qualitätsmanagement oder der lange Atem der Kurzsichtigen

Unter der Etikette “Software-Qualitätsmanagement” werden die Bestrebungen subsumiert, den Prozess der Entwicklung, Auslieferung und Wartung von Software in den Griff zu bekommen, d.h. effizienter zu gestalten. Diese Bestrebungen sind langfristig angelegt. Das Verhalten von Menschen und organisatorische Abläufe sind schwieriger zu ändern als eine Zeile Code. Sie stehen im Wettbewerb um Ressourcen mit der Software-Feuerwehr, die mit dem Lösen der aktuellsten Probleme beschäftigt ist (die vermeintlich einfachere Änderung der Zeilen Code besorgt) – das Kurzfristige behält fast immer die Oberhand. Die Kurzsichtigen haben in den meisten Firmen den längeren Atem als die Weitsichtigen.

Der Teufelskreis ist deutlich erkennbar: Weil keine Prävention gemacht wird, gibt es viele Feuer, die gelöscht werden müssen; weil alle Ressourcen mit Eindämmen der Brandherde beschäftigt sind, gibt es keine freie Ressourcen um zu überlegen, wie man Software-Brände verhüten könnte. Im Themenheft 6/1999 hätten wir gerne Beispiele von Firmen, die den Teufelskreis durchbrochen haben oder gerade intensiv daran arbeiten ihn zu durchbrechen. Ein Teilschritt beim Durchbrechen des Teufelskreises ist das Definieren oder Messen von Produkt-Qualität. Beiträge über praktische Erfahrungen auf diesem Gebiet sind deshalb auch willkommen.

Hinweise für Autoren erhalten Sie auf der Web-Seite http://www.access.ch/sinfo oder auf Anforderung von der Redaktion (nicolet@acm.org). Beiträge sind in *deutscher, französischer und englischer Sprache* willkommen.

Karol Frühauf, Walter Bischofberger, Gastredaktoren
François Louis Nicolet, Redaktor

Termine

Einreicheschluss für die Beiträge	1.8.1999
Benachrichtigung der Autoren	25.8.1999
Einsendeschluss der endgültigen Fassung	20.9.1999
Erscheinungsdatum des Heftes	1.12.1999