September 2001

# Clickstream Warehousing für e-CRM: Neue Herausforderungen an die Datenhaltung?

Ralf Schaarschmidt
*IBM Unternehmensberatung GmbH*, ralf.schaarschmidt@de.ibm.com

Jan Nowitzky
*Friedrich-Schiller-Universität Jena*, nowitzky@informatik.uni-jena.de

Jens Lufter
*Friedrich-Schiller-Universität Jena*, lufter@informatik.uni-jena.de

In: Buhl, Hans Ulrich, u.a. (Hg.) 2001. *Information Age Economy*; 5. Internationale Tagung Wirtschaftsinformatik 2001. Heidelberg: Physica-Verlag

ISBN: 3-7908-1427-X

# Clickstream Warehousing für e-CRM: Neue Herausforderungen an die Datenhaltung?

**Ralf Schaarschmidt**
IBM Unternehmensberatung GmbH

**Jan Nowitzky, Jens Lufter**
Friedrich-Schiller-Universität Jena

*Zusammenfassung: Das vorliegende Papier hat das e-CRM und dessen Datengrundlage zum Thema. Im Web basiert die Beziehung eines Kunden zu einem Unternehmen auf dem Klickverhalten des Kunden. Dieser Clickstream generiert Daten, die in einem speziellen Data Warehouse zur weiteren Analyse gespeichert werden. Die Datenhaltung in einem solchen Clickstream Warehouse muß in die Lage versetzt werden, die ständig wachsenden Datenmengen zu bewältigen und spezifische Anforderungen des Clickstream Warehousing zu erfüllen.*

*Schlüsselworte: CRM, e-CRM, Data Warehouse, Clickstream Warehouse*

## 1 Einleitung

Die verstärkte Kundenorientierung ist heute erklärtes Ziel vieler Unternehmen. Der zunehmenden Konkurrenz und den wachsenden Anforderungen der Kunden soll durch eine an den Kundenbedürfnissen ausgerichtete Verbesserung der Produkte und Dienstleistungen begegnet werden. Unter dem Begriff Customer Relationship Management (CRM) sind alle Maßnahmen zusammengefaßt, die diese kundenzentrierte Ausrichtung von Unternehmen unterstützen [BaÖs00].

Ein weiterer Trend ist die steigende Bedeutung des e-business für die Unternehmen und der damit einhergehende Aufbau technologiebasierter Vertriebskanäle, als wichtigster Vertreter sei das Web genannt. Im Sinne einer umfassenden CRM-Strategie sind alle Vertriebskanäle eines Unternehmens einzubeziehen. Als Teil des CRM beschäftigt sich daher das e-CRM mit Kundenbeziehungen, die sich aus der Interaktion zwischen Kunden und Unternehmen unter Verwendung neuer Technologien ergeben.

Die Interaktion eines Kunden mit einem Unternehmen über das Web wird über das Klickverhalten des Kunden beschrieben. Dieser sogenannte Clickstream ist aus technischer Sicht leicht zu sammeln und stellt eine umfangreiche Datenquelle

zur Beschreibung der Kundenbeziehung dar. Die Voraussetzung für ein erfolgreiches e-CRM ist damit unmittelbar gegeben [MaSp00].

Für die Aufzeichnung von Clickstreams bietet sich ein speziell ausgerichtetes Data Warehouse an. Auf diese Weise steht ein System zur Verfügung, das umfangreiche Analysen der Kundenbeziehung erlaubt. Die Datenhaltung sieht sich jedoch dem Problem der rasant und ständig wachsenden Datenmengen gegenüber. So verzeichnet beispielsweise die bekannte Web-Site Yahoo mehr als 50 Millionen Zugriffe pro Tag.

Das vorliegende Papier wird im weiteren Verlauf einen Bogen spannen, der von der anwendungsbezogenen Aufgabenstellung des e-CRM über deren technologischen Unterstützung bis hin zu den daraus erwachsenden Problemen reicht. In Kapitel 2 werden Begriffe des CRM und e-CRM vorgestellt. Kapitel 3 beschäftigt sich mit den Daten eines Clickstream und deren Speicherung in einem Clickstream Warehouse. Kapitel 4 untersucht die daraus resultierenden Probleme für die Datenhaltung und zeigt Lösungsansätze auf. Kapitel 5 schließt mit einer Zusammenfassung und einem Ausblick.

# 2 e-CRM für Kunden aus dem Web

Im folgenden werden zunächst auf allgemeine Weise Aufgaben und Ziele des CRM diskutiert. Auf dieser Grundlage erfolgt eine kurze Beschreibung der speziellen Aspekte des e-CRM.

## 2.1 CRM

CRM umfaßt alles, was die Beziehung zwischen den Unternehmen und deren Kunden betrifft. Es beinhaltet die Planung, Koordination und Kontrolle aller auf mögliche und bestehende Geschäftsbeziehungen ausgerichteten Aktivitäten. CRM ist damit ein ganzheitlicher Ansatz zur Unternehmensführung. Alle kundenbezogenen Prozesse in Marketing, Vertrieb und Service werden beim CRM bereichsübergreifend integriert und optimiert (Abbildung 1). Die Umsetzung der Kundenorientierung wird anspruchsvoller, da die Kunden für die einzelnen

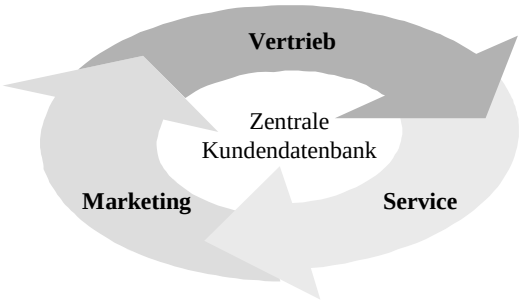


Abbildung 1: Prozesse des CRM

Interaktionsschritte mit dem Unternehmen zunehmend verschiedene Vertriebskanäle (Filiale, Telefon, Post ...) nutzen. Eine datenorientierte Vertriebskanalintegration kann hier Abhilfe leisten.

CRM zielt darauf ab, die Wünsche und Bedarfe von heutigen und zukünftigen Kunden zu verstehen und zu antizipieren. Ziel dieser Strategie ist die profitable Begleitung des Kunden über alle Entwicklungsstufen seiner Beziehung zum Unternehmen: Kundengewinnung, -entwicklung und -pflege.

Die Gewinnung von Kunden ist eine grundsätzliche Aufgabe für ein Unternehmen. Allerdings muß dabei berücksichtigt werden, daß die Akquise von Neukunden drei- bis fünfmal so viele Kosten verursacht wie die Pflege von Bestandskunden. Die Kundenbindung ist daher von besonderer Bedeutung, um die zunehmende Profitabilität einer Kundenbeziehung ausnutzen zu können. Eine Basis für den Erfolg von CRM ist eine Kundensegmentierung im Hinblick auf den Wert eines Kunden für das Unternehmen.

## 2.2 e-CRM als Teil des CRM

Als Teil des CRM wird in diesem Papier das e-CRM verstanden (Abbildung 2). Während sich CRM auf klassische Vertriebskanäle (Filiale, Telefon, Katalog ...) konzentriert, fokussiert e-CRM auf neue, elektronische Vertriebskanäle (Internet, Mobile ...). Die elektronische Quelle Internet kann dabei unmittelbar als wertvolle Informationsquelle über die Kunden und ihr Verhalten dienen. Im Gegensatz zum CRM für die klassischen Vertriebskanäle ist beim e-CRM der Informationsaustausch zwischen Kunde und Unternehmen klar strukturiert. Eine datenbezogene Beschreibung der Kundenbeziehung ist daher direkt herstellbar. Gleichzeitig ist damit zumindest zum Teil die Voraussetzung für die Integration von Vertriebskanälen auf Datenebene geschaffen.

Die aktuell wichtigsten elektronischen Vertriebskanäle sind sicherlich die Web-Sites der Unternehmen. Die browser-basierte Interaktion eines Kunden mit einer Web-Site wird über sein Klickverhalten beschrieben. Die dabei anfallenden Daten werden Clickstream genannt und in einem speziellen Data Warehouse gespeichert.

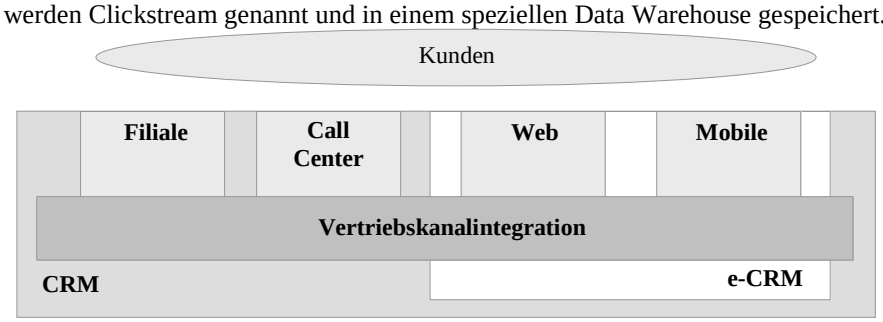


Abbildung 2: e-CRM

# 3 Clickstream und Clickstream Warehouse

Im Gegensatz zur herkömmlichen Gewinnung von Daten über das Kundenverhalten, beispielsweise über Fragebögen, ist ein Clickstream wesentlich umfangreicher und aussagekräftiger aber auch, obwohl strukturiert, unordentlicher. Der reine Clickstream ist jedoch nicht hilfreich, um e-CRM effizient durchzuführen. Hierzu sind eine umfangreiche Aufbereitung und Analyse des Clickstream mit Techniken des Data Warehousing notwendig.

## 3.1 Der Clickstream

Bei der Interaktion von Kunden mit einer Web-Site werden sogenannte Logfiles generiert. Diese Logdaten bilden die technische Basis für den Clickstream.

### 3.1.1 Daten eines Clickstream

Jeder Datensatz im Clickstream stellt einen einzelnen Seitenaufruf eines Nutzers auf einer Web-Site dar. Der Clickstream eines Nutzers ist somit der elektronische Pfad, den der Nutzer beim Navigieren auf einer Web-Site erzeugt. Der Detaillierungsgrad der Aufzeichnung kann dabei sehr hoch sein. Jede aufgerufene Seite, jede benutzte Diskussionsgruppe, der Kauf eines Produktes oder das Kaufinteresse etwa durch Anhören eines Musikstückes wird dabei aufgezeichnet.

Die Aufzeichnung der Interaktion wird vom Web-Server in Form von Logfiles automatisch durchgeführt, d.h. die Daten sind leicht zu sammeln. Auf die unterschiedlichen Log-Formate verschiedener Web-Server soll hier nicht eingegangen

```
141.35.14.100 78612530176216524 – [05/Jan/2001: 21:43:27 +0100]
“GET http://www.google.de/search?q=Clickstream&hl=de&csr=” 2004673
```

Abbildung 3: Beispiel für einen Web-Log-Eintrag

werden, ebenso wenig auf den Umstand, daß i.allg. die Daten eines Clickstream aus mehreren Logfiles verschiedener Web-Server entstehen können. Der Inhalt eines einzelnen Eintrags im Logfile des Web-Servers kann beispielsweise wie in Abbildung 3 dargestellt aussehen.

Diese individuelle Beschreibung des Seitenzugriffs ist wenig aufschlußreich. Aus dem Web-Log werden mittels Transformationen die für das Analysieren von Nutzeraktionen notwendigen Clickstream-Daten erzeugt. Der Clickstream enthält für jede Nutzeraktion folgende Informationen: IP-Adresse und Zeitpunkt des Zugriffs, wenn vorhanden, eine Cookie-ID, die angeforderte Seite bzw. einzelne Objekte dieser Seite, den Typ des Zugriffs, die Seite von welcher referenziert wurde und eine Angabe über den verwendeten Browser.

Abbildung 4 zeigt einen kleinen Ausschnitt aus einem Clickstream. Die aus einem Logfile extrahierten relevanten Nutzeraktionen sind leicht zu interpretieren. Mittels einer Suchmaschine (hier Google mit Suchterm Clickstream) wurde von einem Rechner mit der IP-Adresse 141.35.14.100 auf die entsprechende Startseite zugegriffen. Von dort wurden weitere Aktionen, darunter das Anklicken eines Bildes (logo.gif) und das Navigieren (clickstream-dw.html), ausgeführt. Bei genauerer Betrachtung der Zeiten für die einzelnen Zugriffe fällt auf, daß zwischen

```
141.35.14.100 session of 05/01/01:
21:43:27 /index.html referrer = Google, search = “Clickstream”
21:43:29 /logo.gif
21:43:43 /clickstream-dw.html
21:44:01 /dw-star.html
22:35:01 /index.html
```

Abbildung 4: Ausschnitt aus einem Clickstream

der vorletzten Aktion und der Rückkehr zur Startseite eine sehr lange Zeit verstrichen ist. Eine genaue Aussage, wie lange der Nutzer auf der Seite geblieben ist und ob letztlich dieselbe Person fast eine Stunde später wieder auf die Startseite zugegriffen hat, ist nicht möglich.

Dieses kleine Beispiel macht deutlich, daß die Daten unabhängig davon entstehen, ob der Nutzer ein Kunde ist oder nur ein zufälliger Besucher der Web-Site. Die entscheidende Frage ist nun, wie zum einen einzelne Daten bestimmten Kunden zugeordnet werden können und wie zum anderen alle Daten eines Nutzers sinnvoll zusammengefaßt werden können.

### 3.1.2 Klassifikation der Daten

Von besonderer Bedeutung für das e-CRM sind konkrete Daten über Nutzer. Prinzipiell ist es möglich, die aufgezeichneten Aktionen Nutzern zuzuordnen. Technisch erfolgt dies meist über sogenannte Cookies. Ein Cookie ist eine Technik, mit deren Hilfe ein Web-Server Informationen auf dem Rechner des Nutzers speichert. Die Speicherung von Zuständen und Einstellungen verfolgt bei seriöser Verwendung den Zweck, dem Nutzer beim wiederholten Besuch eine gewohnte und ggf. personalisierte Umgebung anzubieten.

Das Problem aus Sicht einer Web-Site ist jedoch, daß ein Nutzer sich nicht zu erkennen geben muß. Somit lassen sich drei Arten von Nutzern unterscheiden:

- Anonymer Nutzer: Vom Nutzer ist nur eine IP-Adresse bekannt.
- Identifizierter Nutzer: Der Nutzer ist bei wiederkehrendem Besuch über ein Cookie identifizierbar, namentlich aber weiter unbekannt und daher anonym.
- Personifizierter Nutzer: Vom Nutzer sind die persönlichen Daten bekannt.

Im allgemeinen kann der Nutzer einer Web-Site alle Nutzerkategorien durchlaufen. Beim ersten Besuch einer Web-Site ignoriert er vielleicht die Cookies und ist ein anonymer Nutzer des Web-Angebotes. Die IP-Adresse ist zwar bekannt (vgl. Abbildung 4), über diese kann aber auch ein anderer Nutzer auf die Web-Site gelangen, z.B. aufgrund der dynamischen Zuteilung der IP-Adressen durch einen Provider. Im Laufe der Zeit oder aufgrund einer bestimmten Interaktion akzeptiert er ein Cookie (in Abbildung 3 ist “78612530176216524” die Cookie-ID). Jetzt ist der Nutzer identifiziert, aber noch immer unbekannt. Erst wenn er beispielsweise durch den Kauf eines Produktes seine persönlichen Daten hinterläßt, wird aus dem bestimmten, unbekannten Nutzer ein Kunde.

Wichtig aus Sicht des e-CRM sind der identifizierte und der personifizierte Nutzer. Für effektive Analysen können die beiden Benutzergruppen weiter klassifiziert werden, z.B. der Kunde nach seinem Umsatz.

### 3.1.3 Verwendung der Daten

Die Aufgabe des e-CRM besteht darin, durch Integration von Clickstream- und externen Daten (z.B. Produktdaten) kundenzentrierte, vertriebskanalspezifische Analysen durchzuführen, aber auch neue Zusammenhänge zu erkennen [Walt01]. Exemplarisch seien hierfür einige Problemstellungen genannt:

- Identifizierung und Wiedererkennung von Nutzern

  Die Identifizierung eines Nutzers beim Betreten einer Web-Site ist die grundlegendste Aufgabe. Davon ausgehend können gezielte Maßnahmen im Rahmen der Kundengewinnung, -entwicklung und -pflege stattfinden.

- Zielgruppenansprache bei Marketingaktivitäten

  Durch Analyse können umfangreiche Kundenprofile erstellt werden, die letztlich in einer Kundensegmentierung münden. Die unterschiedlichen Kundensegmente orientieren sich dabei i.allg. an der Intensität des Besuchs der Web-Site (neuer Kunde, wiederkehrender Kunde, Vielkäufer etc.).

- Erkennung von abwanderungsgefährdeten Kunden

  Neben der Analyse des aktuellen Kundenverhaltens sind Analysen von lange nicht aktiven Kunden mit dem Ziel der Kundenbindung von Bedeutung. Wichtig hierfür ist das Vorhandensein von Daten über einen langen Zeitraum.

Daneben können auf der Basis der Clickstream-Daten Optimierungspotentiale des Web-Angebots erkannt werden. Hierunter zählen u.a. folgende Aspekte:

- Qualität des Web-Auftritts und der Kundenansprache
- Prüfung des Angebotsportfolios und der angebotenen Dienstleistungen
- Werbeerfolgskontrolle und Messung der Profitabilität
- Bereitstellung von Echtzeitstatusinformation für die Vorgangsbearbeitung

Grundlage für diese Aufgaben bilden umfangreiche Analysen über den zugrundeliegenden Datenbestand bezüglich der Kriterien Zeit und Kunde. Dabei ist es häufig unerheblich, ob alle Nutzer identifizierbar sind. Wichtig ist vielmehr die Tatsache, daß sich verschiedene Nutzer für ein und dasselbe Angebot interessieren. Eine besondere Bedeutung hat dabei die Gewinnung von Kundenprofilen. Die ursprünglichen Daten des Clickstream müssen dabei auf eine aussagekräftige und interpretierbare Ebene abgebildet und aggregiert werden.

## 3.2 Clickstream im Data Warehouse

Für umfassende Auswertungen auch unter Einbeziehung weiterer Datenquellen, für flexible Anfragen und akzeptable Antwortzeiten reichen dateibasierte Analyseansätze nicht aus. Auch im Hinblick auf Skalierbarkeit und Verfügbarkeit ist eine datenbankbasierte Lösung vorzuziehen.

Die Struktur und die Eigenschaften eines Clickstream sowie die anvisierte Verwendung dieser Daten über das Kundenverhalten legen die technische Realisierung in Form eines Data Warehouse nahe [KiMe00]. Das Data Warehouse wird somit zum Clickstream Warehouse und bildet die datenorientierte Basis für ein erfolgreiches e-CRM.

### 3.2.1 Data Warehouse und Data Warehousing

Ein Data Warehouse ist eine themenbezogene, integrierte, zeitbezogene und nichtflüchtige Datenbank zur Entscheidungsunterstützung [Inmo96]. Themenbezogen

heißt hier, daß nicht einzelne Aufrufe von Web-Seiten, sondern Kennzahlen wie Dauer des Besuchs im Mittelpunkt stehen. Die Integration ermöglicht, Daten aus operativen Systemen, wie beispielsweise Kundendaten, zu verwenden. Daten in einem Warehouse werden zeitbezogen erfaßt und spiegeln im Vergleich zu OLTP-Datenbanken nicht nur den aktuellen Stand wider.

Der Begriff Data Warehouse charakterisiert nur die Datenbasis, Data Warehousing umfaßt das gesamte technische Umfeld zur Beschaffung, Speicherung und effizienten Analyse der Daten [BaGü01]. Im Rahmen dieser Arbeit wird die Datenhaltung fokussiert, die vorgelagerte Datenbereitstellung sowie die nachgelagerte Analysephase, Reporting, OLAP, Data Mining (Auffinden von Häufigkeitspunkten oder von Assoziationen im Datenbestand), bleiben außen vor.

### 3.2.2 Star-Schema des Clickstream Warehouse

Ein Clickstream Warehouse ist ein Data Warehouse, welches als wesentlichen Bestandteil den Clickstream von Nutzern einer Web-Site beinhaltet. Durch die Realisierung als Warehouse bildet das Clickstream Warehouse die technische Grundlage für die multidimensionale Datenanalyse, insbesondere der interaktiven, navigatorischen Datenexploration, für das e-CRM. Die zentralen Elemente sind die Kennzahlen (Fakten), die in bezug auf ihre Einflußgrößen angeordnet werden. Kennzahlen im Clickstream Warehouse sind beispielsweise die Dauer des Zugriffs auf eine Web-Seite, die Anzahl der besuchten Web-Seiten in einer Sitzung, die Anzahl der getätigten Aufträge etc.; Einflußgrößen sind z.B. der Kunde, die Zeit, die besuchte Web-Seite, die Herkunfts-Web-Seite und die Sitzung.

Für die Realisierung multidimensionaler Zusammenhänge wird häufig auf relationale Datenbanksysteme zurückgegriffen, um effiziente Analysen bei großen Datenmengen zu gewährleisten. Um multidimensionalen Anfragen („Wieviele Besuche führt ein anonymer Nutzer aus, bevor er sich typischerweise registriert und etwas kauft?“) relational bearbeiten zu können, ist eine Abbildung auf die relationale Ebene notwendig. Varianten zur Umsetzung sind das Star- und das Snowflake-Schema. Beide sind gekennzeichnet durch eine zentrale Faktentabelle, um die verschiedene Dimensionstabellen gruppiert sind (Abbildung 5).

Das Clickstream Warehouse ist durch ein sehr schnelles Wachstum charakterisiert. Verschiedene Varianten der Datenspeicherung sind denkbar:

- Aggregierter Clickstream

  Bevor die Daten in das Warehouse gelangen, werden sie im Hinblick auf die zu erwartenden Analysen aggregiert. Der Clickstream selbst wird nicht weiter verwendet und kann gelöscht werden. Dies schränkt die Auswertemöglichkeiten ein, begrenzt aber die Größe und steigert somit die Handhabbarkeit. Das Clickstream Warehouse wird zum Clickstream Data Mart.

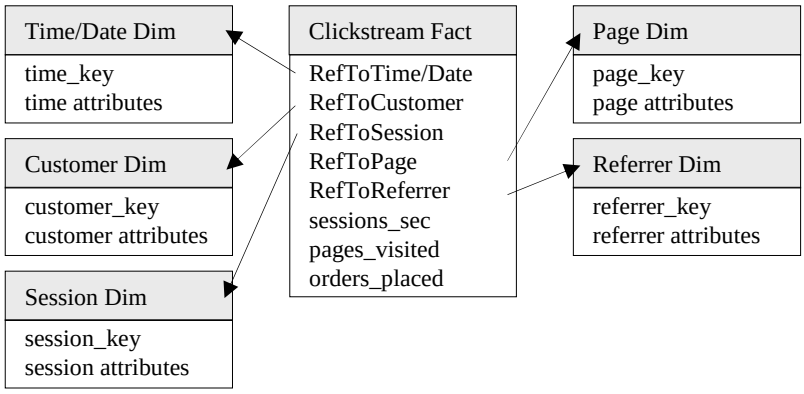


Abbildung 5: Ein Star-Schema für ein Clickstream Warehouse

- Teile des Clickstream

  Vor dem Einspielen der Daten ins Clickstream Warehouse wird der Clickstream von aus Analysesicht unwichtigen Daten befreit, z.B. gelangen Nichtkunden nicht ins Warehouse. Der Clickstream bleibt erhalten und kann für flexible Anfragen genutzt werden. Auch hier werden die Analysemöglichkeiten eingeschränkt, um das Wachstum der Daten zu begrenzen.

- Vollständiger Clickstream

  Der Clickstream wird vollständig ins Warehouse übernommen. Dies beschränkt nicht das Datenvolumen, hält aber im Gegenzug alle Möglichkeiten für Analysen offen.

Vorhandene Daten über Kunden wegzulassen oder so auszudünnen, daß nicht für jeden Kunden die gleichen Analysen möglich sind, schränkt auch die Möglichkeiten des e-CRM drastisch ein und ist daher keine geeignete Strategie. Ein Kompromiß, daß alte Daten gelöscht werden, ist ebenfalls nicht tragfähig. Ein Kunde, der aufgrund seiner getätigten Umsätze ursprünglich als wertvoll galt, über einen längeren Zeitraum aber nicht aktiv war, sollte nicht aus dem Clickstream Warehouse entfernt werden, da es aus Sicht des Unternehmens i.allg. leichter ist, den Kunden zu reaktivieren als einen neuen zu gewinnen. Der Ansatz, daß alle Daten im Clickstream Warehouse verzeichnet sind, wird daher bevorzugt.

Daraus erwachsen allerdings erhöhte Anforderungen an die Datenhaltung. Um ein weiteres Beispiel zu geben: Die Web-Site von Microsoft verzeichnet eine Milliarde Zugriffe am Tag. Dabei erzeugt eine Nutzeraktion einen Log-Eintrag von etwa 200 Byte. Somit wird an einem Tag ein Datenvolumen von 200 Gigabyte erzeugt, welches in der Datenbank zu speichern ist. Dies führt zwangsläufig schnell zu einem Clickstream Warehouse im hohen Terabyte-Bereich.

# 4 Probleme für die Datenhaltung

Im folgenden sollen Probleme und Lösungsansätze diskutiert werden, die sich aus den spezifischen Eigenheiten eines Clickstream Warehouse ergeben.

## 4.1 Anforderungen an ein Clickstream Warehouse

Neben den üblichen Anforderungen an ein Data Warehouse haben sich in den vorherigen Abschnitten folgende Besonderheiten herauskristallisiert:

- Aktualität

  Ein neuer Kunde soll möglichst früh mit personalisierten Angeboten versorgt werden, in der Folge soll die individuelle Kundenansprache auf Basis möglichst aktueller Daten erfolgen. Das Clickstream Warehouse muß also möglichst häufig (z.B. einmal am Tag oder kontinuierlich) mit neu angefallenen Daten aktualisiert werden.

- Feingranularität

  Sinn eines Clickstream Warehouse ist u.a., ein individualisiertes Reagieren zu ermöglichen, das während der Interaktion eines konkreten Kunden mit der Web-Site maßgeschneiderte Daten an diesen zurückliefert. Das impliziert eine sehr große Datenmenge wenig aggregierter Daten.

- eventuelle Echtzeitnutzung

  Neben zeitaufwendigen Analysen auf Basis der gesammelten Fakten ist ein unmittelbares Antwortverhalten auf Aktionen eines Kunden wünschenswert. Die datenbankbasierte Generierung von Web-Seiten erfordert eine extrem schnelle Antwort auf bestimmte Anfragen, da Nutzer einer Web-Site nicht gewillt sind, lange auf das Ergebnis eines Seitenaufrufs zu warten.

Zusammenfassend ergibt sich, daß Clickstream Warehouses sehr groß werden können, häufig aktualisiert werden müssen und ggf. bestimmte Anfragen in Echtzeit zu beantworten haben. Dies führt natürlich zu erhöhten Verfügbarkeitsanforderungen an das Clickstream Warehouse. Kompromisse bezüglich dieser Anforderungen sind möglich, sie gehen aber auf Kosten der Ergebnisgenauigkeit insbesondere bei personalisierten Diensten.

Ein offensichtliches Problem bei der Verarbeitung der durch die Web-Server generierten Daten ist also deren schiere Menge von mehreren Gigabyte pro Tag. Eine intelligente Behandlung der zu bewältigenden Datenmenge durch das dem Clickstream Warehouse zugrundeliegende Datenbanksystem ist daher die Grundvoraussetzung für die Erfüllung unserer Anforderungen, Ansätze dazu sollen im folgenden Abschnitt vorgestellt werden.

## 4.2 Ansätze zum Umgang mit großen Datenbanken

Relationale Datenbanksysteme skalieren bereits heute sehr gut für große Datenmengen, aber im höheren Terabyte-Bereich gibt es noch nicht hinreichend viel Erfahrung. Neue Konzepte zum verteilten Speichern, Abfragen, Sichern und Archivieren derart großer und rasant wachsender Datenbestände sind gefragt: Wann und wie sollen die Daten z.B. aggregiert oder komprimiert werden, welche Indexstrategien werden benötigt, wie müssen die Daten für einen schnellen, parallelen Zugriff verteilt werden?

Wichtige Anbieter von Datenbankmanagementsystemen für den Umgang mit derart großen Datenbanken sind Oracle, IBM (DB2) und NCR (Teradata). Teradata ist dabei spezialisiert auf große Systeme, Oracle und DB2 sind weit verbreitete relationale Datenbankmanagementsysteme, die entsprechend gut skalieren. Nach Angaben der Hersteller liegen selbst Petabyte-Datenbanken noch im Bereich des Möglichen.

Ein aktueller Trend bei den Datenbankanbietern ist die Integration von Funktionen für das Data Warehousing in die Systeme. So bieten die neueren Versionen von DB2 und Oracle z.B. CUBE- und ROLLUP-Funktionen für das einfachere Aggregieren von Warehouse-Daten an, entsprechende Sprachkonstrukte haben auch Aufnahme in die neueste Version der SQL-Norm von 1999 [ISO99] gefunden. Physisch unterstützt wird Data Warehousing in diesen Systemen durch neue Zugriffsvarianten wie Bitmap-Indizes, optimierte Anfragestrategien für Star-Schemata (Star Join) oder verbesserte Möglichkeiten zur Definition und Nutzung materialisierter Sichten [Gupt97].

Mit diesen Techniken lassen sich übliche Data Warehouses heutiger Größenordnungen von fünfzig bis einigen hundert Gigabyte recht gut verwenden. Allerdings muß man verschiedenen Studien nach heute mit einer Verdoppelung der Datenmenge eines typischen Warehouse innerhalb eines Jahres rechnen. Seit 1999 hat sich danach auch die weltweite Zahl der Terabyte-Warehouses verdoppelt, die größten Warehouses umfassen deutlich mehr als zehn Terabyte. Das sind zumeist Installationen großer Firmen der „Old Economy“, einige der größten Warehouses werden z.B. von Wal Mart und UPS betrieben.

Clickstream Warehouses großer Web-Sites holen im Hinblick auf die Datenmenge allerdings schnell auf, weil deutlich mehr Daten anfallen und sie viel einfacher zu sammeln sind. Neben dem Größenwachstum kommen auf Clickstream Warehouses aber auch neue Herausforderungen zu, die sich aus den in Abschnitt 4.1 identifizierten Anforderungen ergeben. So müssen im Prinzip die Analysefunktionen vom Transaktionsbetrieb entkoppelter Data Warehouses mit den Forderungen nach zeitnahen Änderungen und Echtzeitanfragen bei der Online-Nutzung kombiniert werden – OLAP und OLTP treffen sich.

## 4.3 Partitionierung für Clickstream Warehouses

Partitionierung ist eine Form der physischen Verteilung von Daten insbesondere sehr großer Tabellen einer relationalen Datenbank. Ziel ist neben verbesserter Verfügbarkeit und Administrierbarkeit vor allem die Leistungssteigerung bei der Ausführung von Anfragen durch die Einschränkung der zu durchsuchenden Datenmenge und eine Parallelisierbarkeit von Anfragen. Der Einsatz für die riesigen Faktentabellen beim Data Warehousing bietet sich dabei besonders an.

Unter Tabellenpartitionierung [Moha93], kurz Partitionierung, verstehen wir im folgenden die horizontale und vollständige Aufteilung einer Datenbanktabelle in disjunkte Teiltabellen (Partitionen), denen dann gezielt physische Bereiche einer Datenbank zugeordnet werden. Horizontal bedeutet, daß die einzelnen Datensätze selbst nicht zerlegt werden, vollständig und disjunkt besagt, daß jeder Datensatz genau einer Partition zugeordnet wird. Abbildung 6 zeigt beispielhaft eine Partitionierung der Faktentabelle bezüglich zweier Kriterien: Zeit und Kunde. Als Ergebnis erhält man mehrere Partitionen, von denen jede jeweils die Daten einer bestimmten Nutzergruppe eines Jahres enthält.

Es gibt im wesentlichen drei Arten der Partitionierung: eine Verteilung nach dem Round-Robin-Verfahren (reihum ohne weitere Semantik), über eine Hash-Funktion und intervallweise nach den Werten geeigneter Attributkombinationen. Man kann diese Varianten im Rahmen einer mehrstufigen Partitionierung auch miteinander kombinieren [NoMü00].

Während alle drei Partitionierungsarten die Parallelverarbeitung verbessern können, unterstützt insbesondere eine intelligente intervallweise Partitionierung die Optimierung wichtiger Anfragen durch eine dadurch mögliche drastische Einschränkung des Suchraums. Als einfaches Beispiel sei die quartalsweise Aufteilung der Daten einer Tabelle bezüglich eines Datumsattributs genannt. Zeitspezifische Anfragen sind in Data Warehouses häufig. Werden z.B. Angaben benötigt, die sich auf einen Monat beschränken, reicht das Durchsuchen einer einzigen Partition, für ein Geschäftsjahr genügen vier. Gegenüber der über viele Jahre verteilten Daten der Gesamttabelle ist das natürlich eine enorme Ersparnis, die sich unmittelbar in der Leistung niederschlägt.

Für Clickstream Warehouses sind neben zeit- vor allem auch kundenabhängige Analysen interessant, hier kann die Beschleunigung von Anfragen zu Kunden oder Kundengruppen (stärker aggregierte Daten bezüglich bestimmter Kundenprofile) signifikante Auswirkungen im Hinblick auf den gewünschten Echtzeitbetrieb bestimmter Anfragen haben. Entsprechend können wichtige Tabellen der Datenbank bezüglich der Kundendimension, ggf. auch bezüglich Zeit *und* Kunden partitioniert werden (mehrdimensionale Partitionierung), die Partitionierung von Faktentabellen kann auf darüberliegende materialisierte Sichten übertragen werden. Neben den Anfragen ist die Partitionierung auch für das im Vergleich zu herkömmlichen Data Warehouses recht kontinuierliche Einbringen neuer Daten, das bereits

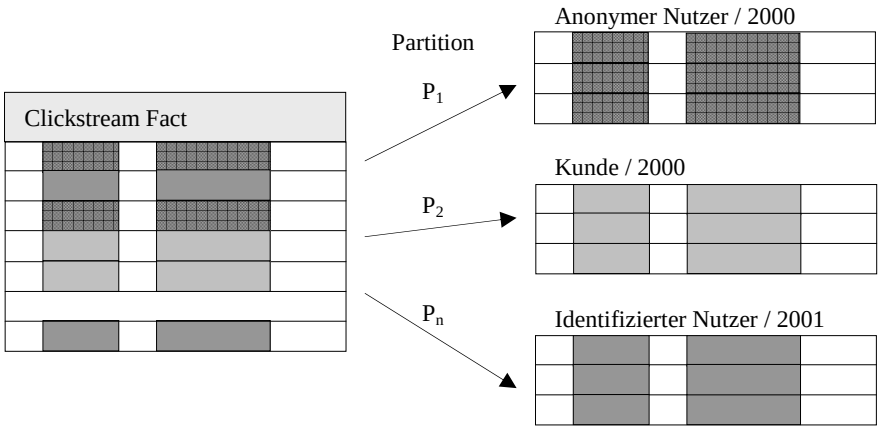


Abbildung 6: Partitionierung einer Clickstream-Faktentabelle

in Abschnitt 4.1 thematisiert wurde, von Interesse. Bei entsprechend gewählten zeitbezogenen Partitionierungskriterien berührt das Einfügen neuer Daten nicht den Anfragebetrieb auf dem Rest der Datenbank.

Die Auswahl geeigneter Partitionierungsstrategien und -kriterien ist nicht trivial, es gibt bislang nur wenige einschlägige Erfahrungen zu komplexeren Anwendungsszenarien. Für weitere Ausführungen sei auf [NoMü00; Nowi00] verwiesen.

# 5 Zusammenfassung und Ausblick

Neue elektronische Vertriebskanäle, insbesondere der Zugriff über das Web, erweitern derzeit das klassische Customer Relationship Management um zusätzliche Varianten der Interaktion mit Kunden. Die dabei anfallenden Daten sind eine wertvolle Informationsquelle über Kunden und deren Verhalten und sollen im Rahmen des e-CRM möglichst effektiv genutzt werden.

In diesem Papier wurde dazu der Ansatz des Clickstream Warehousing vorgestellt, bei dem die während der Interaktion eines Nutzers mit der Web-Site eines Unternehmens anfallenden Daten (Clickstream) in ein Data Warehouse eingebracht werden. Im Vergleich zum traditionellen Data Warehousing ergeben sich zusätzliche Anforderungen wie Aktualität, geringe Aggregierung und Echtzeitnutzung des Warehouse, die zusammen mit der inhärent großen und schnell wachsenden Datenmenge neue Herausforderungen an die Datenhaltung stellen.

Führende Hersteller von Datenbanksystemen haben ihre Produkte in den vergangenen Jahren um wichtige Funktionen für das Data Warehousing erweitert, weitere Verbesserungen z.B. zur Echtzeitnutzung sind geplant. Die Handhabung sehr großer Data Warehouses im Terabyte-Bereich ist allerdings noch längst kein All-

gemeingut. Um mit dem Größenwachstum insbesondere von Clickstream Warehouses Schritt zu halten, sind zusätzliche Ideen ständig gefragt.

Ein hier vorgestellter relativ neuer Ansatz zur Bewältigung dieser Aufgabe ist eine Partitionierung von Tabellen im Warehouse, die zu deutlichen Leistungsverbesserungen insbesondere bei der Anfrageverarbeitung führen kann. Ein Schwerpunkt für weitere Arbeiten an der Universität Jena liegt unter anderem in der Auswahl geeigneter Partitionierungsstrategien für den Warehouse-Bereich und deren physische Unterstützung, hier gibt es bislang nur wenige Erfahrungen.

# Literatur

[BaGü01] A. Bauer und H. Günzel. Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung. dpunkt-Verlag, Heidelberg, 2001.

[BaÖs00] V. Bach und H. Österle (Hrsg.). Customer Relationship Management in der Praxis: Erfolgreiche Wege zu kundenzentrierten Lösungen. Springer-Verlag, Berlin, Heidelberg, 2000.

[Gupt97] H. Gupta. Selection of views to materialize in a data warehouse. In Proceedings of the 6th Int. Conference on Database Theory, Seiten 98-112, Delphi, Greece, Januar 1997.

[Inmo96] W.H. Inmon. Building the Data Warehouse. John Wiley & Sons, New York, 1996.

[ISO99] ISO/IEC 9075-2:1999. Information Technology – Database Languages – SQL – Part 2: Foundation (SQL/Foundation), 1999.

[KiMe00] R. Kimball und R. Merz. The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse. John Wiley & Sons, New York, 2000.

[MaSp00] B.M. Masand und M. Spiliopoulou (Hrsg.). Web Usage Analysis and User Profiling. In Proceedings of the Int. WEBKDD'99 Workshop, San Diego, CA, August 1999, LNCS, Vol. 1836, Springer-Verlag, 2000.

[Moha93] C. Mohan. A survey of DBMS research issues in supporting very large tables. In Proceedings of the 4th Int. Conference on Foundations of Data Organization and Algorithms, Seiten 279-300, Chicago, IL, Oktober 1993.

[NoMü00] J. Nowitzky und T. Müller. Entwurf und Bewertung von Partitionierungsstrategien für Datenbankschemata. Jenaer Schriften zur Mathematik und Informatik. Math/Inf/00/29, Institut für Informatik, Friedrich-Schiller-Universität Jena, Oktober 2000.

[Nowi00] J. Nowitzky. Tabellenpartitionierung für die Archivierung im SAP System R/3. Jenaer Schriften zur Mathematik und Informatik. Math/Inf/00/06, Institut für Informatik, Friedrich-Schiller-Universität Jena, März 2000.

[Walt01] R. Walther. Web Mining (Aktuelles Schlagwort). Informatik Spektrum 24(1):16-18, Februar 2001.