Business Intelligence – Data Warehouse für den Ferienclub

Jan Weinschenker

8. Juli 2005

Im Rahmen der Vortragsreihe im Fach Anwendungen I beschäftigt sich diese Ausarbeitung mit dem Thema *Data Warehousing*.

Informationssysteme in Unternehmen, aber auch in Behörden und Forschungsinstitutionen sind in den vergangenen Jahren zunehmend komplexer geworden. Häufig kann eine einzelne Person die ablaufenden Prozesse und die angefallenen Daten nicht mehr alleine überblicken. Die sich im Einsatz befindlichen Informationssysteme sind in aller Regel von höchst heterogener Natur. Man hat es nicht mit Systemen „aus einem Guß“ zu tun, sondern mit einem bunten Zoo von Anwendungen, die ursprünglich als „Stand-Alone“-Anwendung entworfen wurden. Eine Integration der Systeme schreitet oft nur langsam voran. Weiterhin liegen die gespeicherten Daten oft in einer Form vor, die nicht für eine tiefgreifende Analyse geeignet ist. Sie sind in der Regel anwendungs- und nicht analyseorientiert modelliert worden.

Als Beispiel können hier die Informationssysteme eines typischen deutschen Versicherungsunternehmens betrachtet werden. Dort gibt es mehrere Sparten, wie zum Beispiel Krankenversicherung, Kraftfahrzeuge, Reise- und Lebensversicherung. Mit der Einführung moderner IT schufen sich die Sparten zunächst komplett eigene Informationssysteme. Die verwendeten Datenmodelle sind sehr unterschiedlich. Bei der Krankenversicherung sind Geschlecht, Alter und Vorerkrankungen einer Person von Bedeutung. Über die Laufzeit eines Vertrages kommt es relativ häufig vor, dass Leistungen beansprucht werden, nämlich immer dann, wenn der Versicherte einen Arzt besucht. Bei einer Lebensversicherung werden zwar noch ähnliche personenbezogene Daten gespeichert, die Leistungsabwicklung unterscheidet sich jedoch erheblich. Noch größer sind die Unterschiede bei den anderen genannten Sparten. Beim Entwurf dieser Systeme war, aufgrund der unterschiedlichen Datenmodelle und der damaligen Technik, oft keine Integration der Daten vorgesehen.

Zur Vorbereitung Managemententscheidungen ist es jedoch notwendig, dass Analysen über alle vorhandenen Daten ermöglicht werden. Grundlage dafür ist eine einheitliche Sicht auf den gesamten Datenbestand eines Unternehmens. Mit einem Data Warehouse wird nachfolgend ein Werkzeug vorgestellt, das diese Analysen ermöglichen kann.

In einem kurzen Grundlagenteil wird zunächst die prinzipielle Funktionsweise eines Data Warehouse erläutert. Anschließend wird die Referenzarchitektur für Data Warehouses nach [BG+04] vorgestellt. Zum Schluss wird darauf eingegangen, welchen Beitrag ein Data Warehouse für das aktuelle Projekt *Ferienclub* im Master-Programm für verteilte Systeme leisten kann.

Für tiefer gehende Informationen zum Thema Data Warehouses empfehle ich die Lektüre von [HW05], [Inm96] und [BG+04]. Insbesondere die letztgenannte Quelle war auch die Grundlage dieser Ausarbeitung.

Inhaltsverzeichnis

1 Grundlagen

Dieser Abschnitt geht kurz auf die Motivation ein, die allgemein hin hinter der Einführung von Data Warehouses steht. In einem weiteren Unterabschnitt werden gängige Verfahren des Data Warehousing erläutert. Der Schwerpunkt liegt dabei auf dem Import von Anwendungsdaten in das Data Warehouse und die Aufbereitung dieser Daten für eine anschließende Analyse. Analyseverfahren selbst sind nicht Gegenstand dieser Ausarbeitung, siehe dazu [Elv05].

1.1 Definition und Motivation

In der Einleitung wurde bereits ein Beispiel erläutert. Die Kernidee ist, dass aus den unterschiedlichen Datenquellen, die in einem Unternehmen, einer Behörde oder einer anderen Institution existieren, eine einheitliche Datenbasis erstellt wird. Diese Datenbasis wird Grundlage für Analysevorgänge. Entsprechend definiert [Inm96] auch den Begriff *Data Warehouse*:

> A data warehouse is a subject oriented, integrated, non-volatile and time variant collection of data in support of management‘s decisions.

Hier werden noch zwei weitere Aspekte angesprochen. Zum einen müssen Auswertungen über die Zeit möglich sein. Die historische Entwicklung von Sachverhalten soll dokumentiert werden

können. Weiterhin besteht die Anforderung, dass Werte innerhalb des Data Warehouse nicht mehr verändert werden dürfen.

Die Anwendungsfelder liegen jedoch nicht nur im wirtschaftlichen Bereich. Überall dort, wo große Datenmengen integriert und analysiert werden, können Data Warehouses eine große Hilfe sein. Aus diesem Grund beschränken sich [BG+04] in ihrer Definition nicht auf einen bestimmten Anwendungsbereich:

> Ein Data Warehouse ist eine physikalische Datenbank, die eine integrierte Sicht auf beliebige Daten zu Analysezwecken ermöglicht.

Hier wird der Schwerpunkt eher darauf gelegt, dass das Data Warehouse eine zentrale Instanz darstellt. Auch wenn die Quelldaten der Produktionssysteme ursprünglich verteilt vorliegen, so soll das Data Warehouse diese in einer vereinheitlichten Sicht wiedergeben.

Weiterhin hat letztere Definition keine Einschränkung hinsichtlich des Arbeitsumfeldes. Es sollen nicht nur Managemententscheidungen unterstützt werden. Dies ist mittlerweile näher an der Realität. Insbesondere in wissenschaftlichen oder technisch ausgerichteten Umgebungen fallen große Datenmengen an, die mit einem Data Warehouse bequem analysiert werden können.

1.2 Verfahren

Im Folgenden werden einige Verfahren des Data Warehousing erläutert, die im Rahmen des Projekts *Ferienklub* von Bedeutung sind. Die Ausführungen konzentrieren sich dabei auf den Import von Daten in das Data Warehouse und deren Transformation in eine analysetaugliche Form. Diese Vorgänge werden in der Literatur oft unter den Schlagworten Extraktion, Tranformation, Lader oder kurz ETL zusammengefasst. Für weitere Ausführungen, die sich mit der eigentlichen Analyse beschäftigen, siehe [Elv05]. Viele hier vorgestellt Sachverhalte setzen Grundwissen im Bereich von Datenbanken voraus. Für weitere Informationen zu diesem Thema empfehle ich [DB].

1.2.1 Extraktion

Wie bereits erwähnt, soll ein Data Warehouse an einer zentralen Stelle eine einheitliche Sicht auf einen Datenbestand gewährleisten. Zu diesem Zweck müssen die gewünschten Daten zunächst aus den Produktivsystemen in das Data Warehouse importiert werden.

Dieser Vorgang wird als Extraktion bezeichnet. Die Art der Extraktion richtet sich nach der Art der Datenquelle. Liegen die Daten in einer relationalen Datenbank (RDBMS), wird häufig über sogenannte Bulk-Load-Werkzeuge importiert. Ein Bulk-Loader greift nicht über die in RDBMS übliche Anfragesprache SQL auf die Daten zu, sondern kann die Daten Tabellenweise in großen Mengen transferieren.

Es gibt jedoch Fälle, in denen die Quelldaten in anderen Formaten vorliegen. Beispiele wären hier Excel-Tabellen oder Access-Datenbanken, hierarchische oder objektrelationale Datenbanken. In diesen Fällen müssen spezielle Extraktions-Werkzeuge zum Einsatz kommen, mit einer Vielzahl von Schnittstellen ausgestattet sind.

Ein weiter wichtiger Aspekt ist der Zeitpunkt der Extraktion, also die Frage, wann Extraktionen erfolgen sollen oder müssen. Werden große Datenmengen extrahiert, so kann dadurch die Performance des Quellsystems erheblich beeinflusst werden. Dieser Zustand sollte natürlich möglichst vermieden werden. Man führt deswegen diese Extraktionen zu Zeiten durch, in denen wenig oder garnicht mit dem Produktivsystem geareitet wird. Probleme ergeben sich, sobald man mit Systemen zu tun hat, die rund um die Uhr genutzt werden. Wird ein System weltweit genutzt, existiert immer eine Zeitzone, in der gerade normale *Office hours* sind.

Anforderungen, die durch die Analyse gestellt werden, nehmen ebenfalls Einfluss auf den Extraktionszeitpunkt. So ist es möglich, Daten nur bei Bedarf in das Data Warehouse zu importieren. Die Extraktion erfolgt also zu dem Zeitpunkt, wo eine bestimmte Anlayse erstellt werden soll. Eine weitere Möglichkeit ist, dass die Produktivsysteme von sich aus Daten exportieren, sobald eine Änderung stattgefunden hat. In modernen DBMS können Trigger-Mechanismen auf solche Änderungen reagieren und von sich aus die Extraktion starten.

1.2.2 Transformation

Die Extraktionskomponente transferiert die gewünschten Daten in den Arbeitsbereich. Auf diese Komponente des Data Warehouse wird in einem späteren Abschnitt genauer erläutert. Im Arbeitsbereich erfolgt die Transformation der Daten in eine analyseorientierte Form.

Daten aus unterschiedlichen Quellen werden sinnvoll zusammengefasst. Ein großes Problem dabei ist, dass je nach Quelle ein anderes Datenformat vorliegt. Es kann beispielsweise vorkommen, dass Postleitzahlen in einem System als String gespeichert werden und in einem anderen als Zahlenwert. Der Wert kann in einem System, wie erwartet, *Postleitzahl* heißen. Woanders trägt er vielleicht die Bezeichnung *ZIP-Code*. Ziel der Transformation muss es jetzt sein, alle Daten so umzuwandeln, dass man einheitlich mit ihnen arbeiten kann. Hierzu ist Domänenwissen über die einzelnen Quellsysteme notwendig.

Werden zeitbezogene Daten zusammengefasst, müssen eventuell die Unterschiede zwischen mehreren Zeitzonen vereinheitlicht werden, d. h. alle Angaben werden auf eine bestimmte Zeitzone umgerechnet. Bei Geldbeträgen, die in den Quellsystemen in unterschiedlichen Währungen gespeichert werden, ist möglicherweise die Umrechnung in eine einheitliche Währung von Nöten.

Ein anderes Problem ergibt sich, wenn sich Daten aus den Quellsystemen gegenseitig widersprechen. Soll zum Beispiel auf Kundendaten zugegriffen werden, kann es vorkommen, dass für die selbe Person unterschiedliche Adresse oder Geburtsdaten vorliegen. Es ist schwierig, diese Widersprüche automatisch aufzulösen. Im einfachsten Fall entscheidet man sich, generell den Daten aus einem bestimmten Quellsystem den Vorzug zu geben. Im schlimmsten Fall müssen Widersprüche manuell aufgelöst werden, d. h. jemand muss für jeden Fall entscheiden, wie die Daten übernommen und zusammengefasst werden.

1.2.3 Laden

Der Ladeprozess innerhalb eines Data Warehouse kommt in mehreren Ausprägungen zum Einsatz. Das Laden ist der rein technische Transfer von Daten zwischen den Komponenten des Data Warehouse. Da bei diesen Transfers oft sehr große Datenmengen bewegt werden, ist hier die Effiziens und die Performance von großer Bedeutung. Wie schon bei der Extraktion kommen auch hier Bulk-Load-Werkzeuge zum Einsatz, die die herkömmlichen Datenbankschnittstellen umgehen können.

Von großer Bedeutung beim Laden ist eine zuverlässige und ausgeklügelte Fehlererkennung- und behandlung. Insbesondere ist es nicht erlaubt, bestehende Datensätze zu ändern oder diese gar zu überschreiben. Um dies zu verhindern werden die Daten normalerweise mit einer Zeitdimension versehen und die neuen Daten werden als „aktueller Stand“ zum existierenden Bestand hinzugefügt.

2 Referenzarchitektur

In [BG+04] wird eine modulare Referenzarchitektur vorgestellt. Anhand der einzelnen Komponenten kann die funktionsweise des Data Warehouse sehr anschaulich erläutert werden. Abbildung 1 auf Seite 5 zeigt eine vereinfachte Version der Referenzarchitektur.

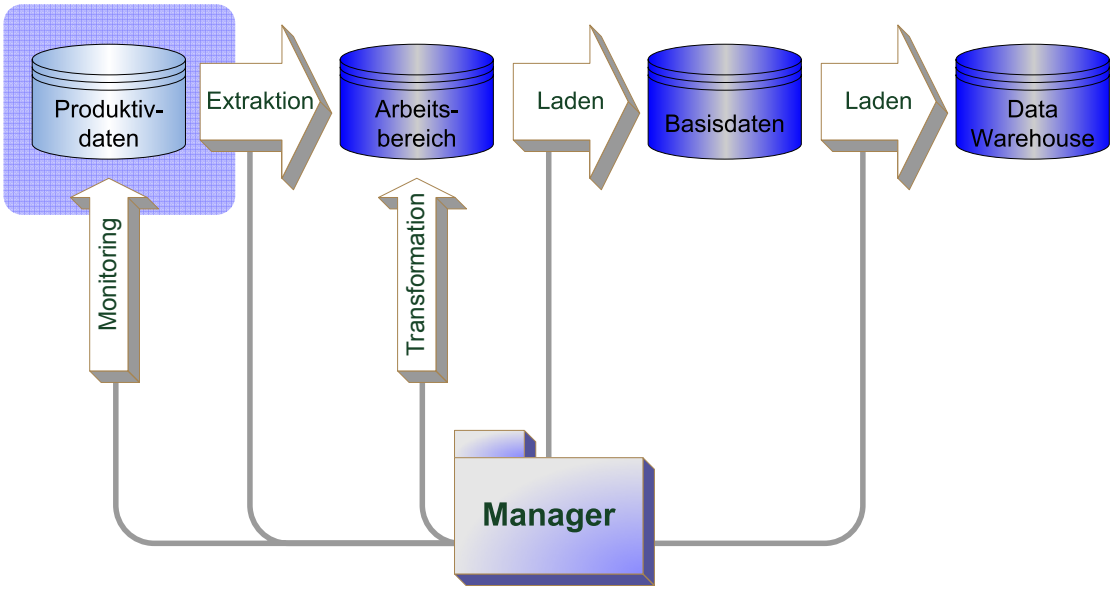

Abbildung 1: Vereinfachte Referenzarchitektur nach [BG+04]

2.1 Datenhaltung

Es existieren vier Komponenten der Datenhaltung, die im Folgenden näher erläutert werden.

2.1.1 Produktivdaten

Die Produktivdaten stehen stellvertretend für die externen Datenquellen. Als solche gehören sie nicht direkt zum Data Warehouse, sind jedoch für die Architektur relevant. Dort liegen die Daten in anwendungsorientierter Form vor und werden von dort durch den Extraktionsprozess importiert.

2.1.2 Arbeitsbereich

An diesem Ort landen die importierten Daten zuerst, nachdem sie aus den Produktivsystemen extrahiert wurden. Hier werden sie transformiert, integriert und fehlerbereinigt (siehe Abschnitt 1.2.2). Anschließend werden die Daten in die Basisdatenbank geladen und aus dem Arbeitsbereich gelöscht. Sie sind dort also nur temporär vorhanden.

2.1.3 Basisdatenbank

Die Basisdatenbank ist der zentrale Aufbewahrungsort der Daten im Data Warehouse. Hier liegen sie bereinigt, aufbereitet und in der kleinsten Granularität vor, sind allerdings noch nicht analyseorientiert. An diesem Punkt hat noch keine Aggregation oder sonstige Zusammenfassung der Daten stattgefunden. Bei diesen Verfahren, die während der Analyse angewandt werden, gehen oft auch wieder Informationen verloren. Die Basisdatenbank soll jedoch Grundlage für

alle weiteren Analysen sein, weswegen auch Änderungsoperationen auf den Daten, im Gegensatz zum Arbeitsbereich, nicht mehr zulässig sind.

2.1.4 Data Warehouse

Das eigentliche Data Warehouse als Bestandteil eines Data Waehouse Systems ist der Ort, an dem die Analysen durchgeführt werden. Das Data Warehouse wird mit den benötigten Daten aus der Basisdatenbank gefüllt. Mehr zu den Vorgängen und Verfahren, die ab hier zum Tragen kommen, ist nachzulesen in [HW05].

2.2 Prozesse und Metadaten

Dieser Abschnitt geht noch einmal kurz auf die Prozesse ein, die innerhalb eines Data Warehouse ablaufen. Die zum ETL gehörenden Vorgänge wurden bereits in Abschnitt 1.2 erläutert und tauchen hier nicht noch einmal auf. Stattdessen werden das Monitoring und die Metadaten vorgestellt.

2.2.1 Monitor

Beim Monitoring werden Änderungen an den Produktivdaten überwacht. Ein Prozess, der Monitor, erkennt Änderungen an einer Datentabelle und meldet diese dem Data Warehouse Manager (siehe 2.3, der daraufhin eine Extraktion initiiert.

Änderungen können erkannt werden, indem beispielsweise der Zustand einer Datentabelle zu einem Zeitpunkt x mit dem Zustand zum Zeitpunkt $x + 1$ verglichen wird. Bei einer Differenz liegt eine Änderung vor. Der Abstand zwischen den zwei Zeitpunkten hängt von den Analyseanforderungen und der Natur der Quelldaten ab. Es wurde bereits erwähnt, dass die Quellsysteme von sich aus eine Änderung an das Data Warehouse, oder genauer, den Monitor melden können.

2.2.2 Metadaten

In den Metadaten steckt das Wissen des Data Warehouse über die Strukturen der Quell- und Basisdaten sowie der Data Warehouses. Mit diesen Informationen werden alle ablaufenden Prozesse und damit der Datenfluss gesteuert. Die Extraktion erfährt aus den Metadaten, welche Daten von wo importiert werden müssen. Auch die genaue Art und Weise der Transformation kann durch Metadaten beschrieben werden.

In der dargestellten Architektur wird das Idealbild einer einzigen Metadatenbank dargestellt. In [BG+04] wird das Problem beschrieben, dass viele Komponenten eines Data Warehouse in der Praxis häufig eigene Metadaten verwenden, und das in diesem Bereich keine Integration stattfindet. Die Metadaten der Extraktion sind also getrennt von denen der Transformation oder der Lade-Prozesse.

2.3 Data Warehouse Manager

Der Manager ist die steuernde Komponente des Data Warehouse. Er initiiert, steuert und überwacht alle ablaufenden Prozesse. Insbesondere die Initiierung der Datenschaffungsprozesse, also des ETL, stellt seine Hauptaufgabe dar. Dabei ist ebenfalls wichtig, dass die Vorgänge protokolliert werden. Die Herkunft der Daten muss nachvollziehbar sein. Weiterhin muss eine Fehler- und Ausnahmebehandlung möglich sein. Der Manager sollte mit Problemen in einem gewissen Umfang eigenständig umgehen können oder bei Bedarf einen Administrator verständigen.

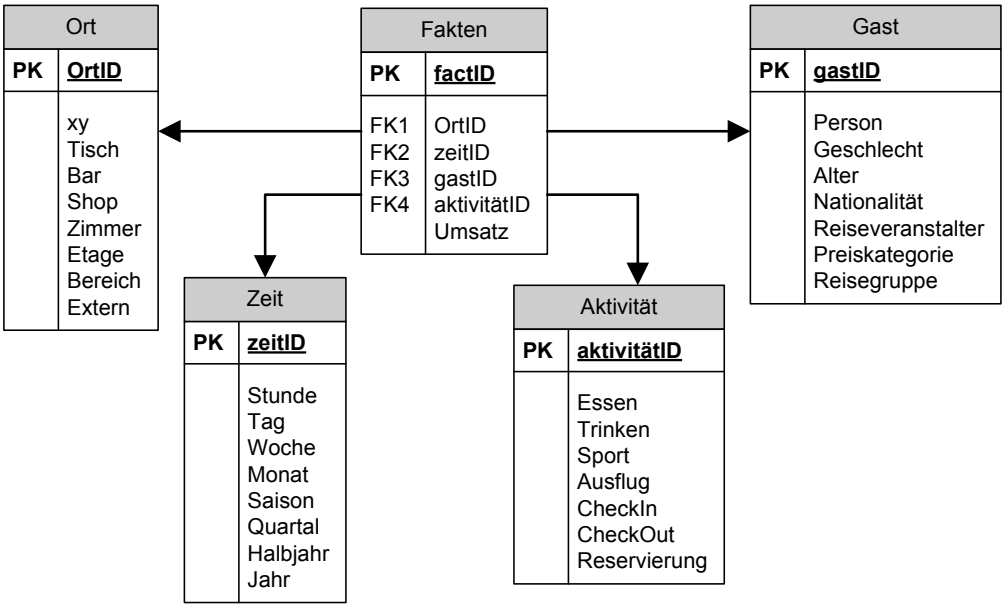

Abbildung 2: Starschema: Vorschlag für ein Datenmodell

3 Anwendungsmöglichkeiten im Projekt Ferienclub

Das Projktziel ist, unter der Metapher eines Ferienclubs verschiedenste Technologien und Konzepte aus aktuellen Themenfeldern der Informatik anzuwenden und mehr oder weniger stark miteinander zu integrieren. Da das Projekt im Rahmen des Master-Programms *Verteilte Systeme* stattfindet und viele Komponenten des Ferienclubs verteilt arbeiten werden, hat es sich die Teilgruppe Business Intelligence zum Ziel gemacht, für eine einheitliche Sicht auf die anfallenden Daten zu sorgen.

Abbildung 2 auf Seite 7 zeigt einen Vorschlag für ein Datenmodell des Data Warehouses. Da es ein Teilprojekte geben wird, welche RFID oder Mobile Computer einsetzen, wird es möglich sein, ortsbezogene Daten zu sammeln. Weiterhin wird es eine dienstorientiere Infrastruktur (SOA) geben, mit der eventuell Web-Services oder ähnliche Dienste und Anwendungen realisiert werden können. Die damit gesammelten Daten können mit einem Data Warehouse integriert und Analysiert werden.

Beim Entwurf des abgebildeten Beispiels wurde unter anderem daran gedacht, dass es möglich sein wird, von einem PDA aus Veranstaltungen oder Mietwagen zu buchen. Möglich wäre auch Essen oder Getränke darüber zu bestellen oder Auskunftsdienste zu nutzen. Über die damit gesammelten Daten ließen sich Benutzer- und Interessenprofile entwickeln, die wiederum anderen Diensten zur Verfügung gestellt werden könnten. Für weitere Ausführungen zu den Analysemöglichkeiten, siehe [Elv05].

A Glossar

DBMS Datenbank Management System

ETL Abkürzung für Extraktion, Tranformation, Laden. Siehe [BG+04].

PDA Personal Digital Assistant. Ein tragbarer Kleincomputer, der sogar über eine Mobil- oder Drahtlosnetzwerkanbindung verfügen kann

RDBMS Relationales Datenbank Management System

RFID Radio Frequency Identification

SOA Service Orientiented Architecture

SQL Structured Query Language – eine Anfragesprache für RDBMS

B Literatur

[BG+04] Bauer, Andreas, Holger Günzel et al.: *Data Warehouse Systeme – Architektur, Entwicklung, Anwendung.* dpunkt.verlag, 2. Auflage, 2004.

[Elv05] Elvers, Sven: *Business Intelligence – Datamining und Datenexport.* Hochschule für Angewandte Wissenschaften Hamburg, Master-Programm Verteilte Systeme, Juni 2005.

[HW05] Humm, Bernhard und Frank Wietek: *Architektur von Data Warehouses und Business Intelligence Systemen.* Informatik Spektrum, Band 28 Heft 1:3–14, Februar 2005.

[Inm96] Inmon, W. H.: *Building the Data Warehouse.* John Wiley & Sons, New York, 2. Auflage, 1996.