

Business Intelligence

Data Warehouse / Analyse

Sven Elvers

2005-07-06

Einleitung

Dieses Dokument beschreibt einen für das Verständnis relevanten Teil der Präsentation.

Business Intelligence

„...Allgemein umfasst der Begriff die analytischen Konzepte, Prozesse und Werkzeuge, um Unternehmens- und Wettbewerbsdaten in entscheidungsrelevantes Wissen zu transformieren. Es werden unternehmensinterne und -externe Daten als Quellen herangezogen...“ [WIKI 05]

Motivation

Business Intelligence wird eingesetzt, um vorhandene Information aufzubereiten und in einer verständlicheren Form darzustellen. Es wird aber auch dazu verwendet, um aus den Informationen neues Wissen zu erzeugen oder um die betriebsweiten Daten in entscheidungsrelevantes Wissen zu transformieren, um Entscheidungen zu unterstützen.

Referenzarchitektur

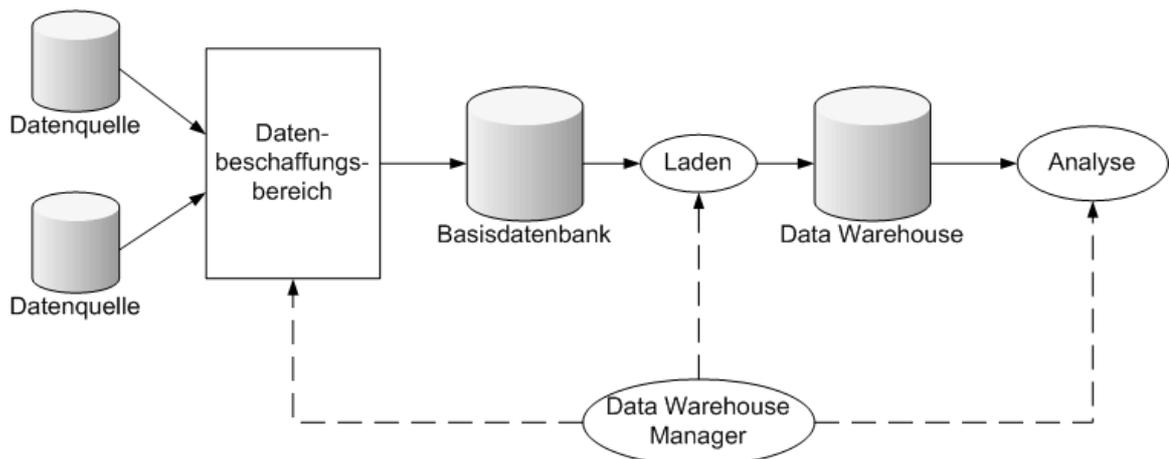


Abb. 1 Referenzarchitektur

Abb. 1 zeigt die Referenzarchitektur in vereinfachter Darstellung, da Jan Weinschenker sich um den Datenbeschaffungsbereich kümmern wird (siehe [ANW1]). Ich werde im Weiteren kurz die Komponenten Basisdatenbank, das allgemeine Datenlager, und Data Warehouse, das für spezielle Analysen ausgerichtete Datenlager, erläutern, welche die Schnittstellen zwischen Jan Weinschenkers und meinem Beitrag sein werden. Danach werde ich etwas ausführlicher auf die Analyse-Komponente eingehen, da diese der Hauptaspekt meines Projektbeitrages sein wird.

Basisdatenbank

Die Basisdatenbank bietet eine integrierte Sicht auf die Datenquellen mit bereinigten Daten. Sie ist umfassend bezüglich Zeit und Granularität, da noch keine Aggregationen ausgeführt wurden und sie nicht analysebezogen aufgebaut ist, sondern nur die Daten aus der Produktionsumgebung sammelt.

Durch den Einsatz der Basisdatenbank gibt es eine gemeinsame Schnittstelle für alle Datenquellen und Data Warehouses (Narbe-Speicher-Architektur), wodurch die Anzahl

der Schnittstellen von $m \cdot n$ (Anzahl Quellen mal Anzahl Data Warehouses) auf $m+n$ reduziert wird. Dadurch wird das System flexibler, da z.B. die Datenbasis für neue Analysen bereits vorhanden ist.

Data Warehouse

Für das Data Warehouse werden ausgewählte Daten aus der Basisdatenbank verdichtet in das mehrdimensionale Datenmodell des Data Warehouse geladen. Dieses Modell ist speziell für die Analysen ausgerichtet. Es besteht aus Fakten (z.B. Umsatz) und Dimensionen (z.B. Produkt, Zeit, Filiale).

Das Laden der Daten aus der Basisdatenbank ins Data Warehouse wird über ein „bulk loader“ realisiert, da davon ausgegangen wird, dass die Basisdatenbank immer konsistent ist und auf dem Data Warehouse keine Schreiboperationen ausgeführt werden, außer vom „bulk loader“. Der „bulk loader“ ist ein Tool, um schnell große Datenmengen in eine Datenbank zu spielen, indem die Mehrbenutzerkoordination, die Konsistenzprüfung etc. umgangen werden.

Das Data Warehouse kann aus so genannten einzelnen „data marts“ bestehen. Dieses wird praktiziert, um unter anderem die Last zu verteilen oder einen Performanzgewinn durch Aggregation zu erhalten.

Analyse

Bei den Analyseverfahren werden drei Hauptkategorien unterschieden:

- Data Access
- Online Analytical Processing (OLAP)
- Data Mining

Unter *Data Access* gehören die einfachen Datenbankabfragen. Bei jeder Anfrage muss ein neues Statement erstellt werden. Die Reporting Tools verwenden diese Art der Analyse, da zumeist festgelegt ist wie der Report auszusehen hat.

Bei Tools die *OLAP* unterstützen braucht der Benutzer nur einmal am Anfang alle Kriterien zu definieren und kann sich dann in dem Dimensionswürfel des Data Marts bewegen. Mit Pivotierung kann er die Reihenfolge von Spalten, Zeilen, etc. ändern und mit Rotation kann er den Blickwinkel auf den aktuellen Ausschnitt des Würfels ändern. Roll-up und Drill-down werden dazu verwendet, um die Granularität zu verändern. Wobei Roll-up das angezeigte Ergebnis grob-granularer macht, z.B. der Wechsel der Filialsicht von der Bezirksebene auf die Ländersicht, und Drill-down lässt das Ergebnis fein-granularer werden. Bei Drill-across wird der betrachtete Würfel gewechselt, z.B. werden danach die Umsätze anstelle der Verkäufe angezeigt. Das Einschränken der Information von einer Ebene (Dimension) wird Slice genannt und Dice ist die Einschränkung von mehreren Ebenen (Dimensionen).

Die Analyseverfahren, die unter *Data Mining* zusammengefasst sind, versuchen aus den vorhandenen Informationen neue Informationen zu gewinnen. Beim Clustering werden die vorhandenen Informationen in mögliche Gruppen (Cluster) zusammengefasst und bei Klassifikation werden die Informationen mit einer bestimmten Wahrscheinlichkeit einer festen Klasse zugeordnet. Zusammenhänge zwischen Ereignissen und ihren Auswirkungen werden mit den Regressionsanalysen ermittelt. Bei den Assoziationsanalysen wird untersucht, welche Beziehungen zwischen einzelnen Merkmalsausprägungen bestehen und die Abweichungsanalyse hat das Ziel

Ausprägungen von Merkmalen zu entdecken, die sich stark von den anderen Merkmalsausprägungen unterscheiden.

Die Warenkorbanalyse ist ein Beispiel für Assoziationsanalysen, bei der zuerst bestimmt wird wie häufig ein Produkt in den Warenkörben vorkommt (support). Dann wird durch ein festgelegtes minsupp bestimmt, welche Produkte in die Berechnung übernommen werden. Jetzt werden Regeln in der Form $A \rightarrow B$ (Wenn Produktmenge A im Korb ist, dann ist auch Produktmenge B im Korb) aufgestellt, wobei B und A keine gemeinsamen Elemente enthalten. Für diese Regeln werden dann die Wahrscheinlichkeiten (confidence) berechnet, wenn die Bedingung (A) wahr ist, auch die jeweilige Regel wahr ist. Anhand des minconf wird danach bestimmt, welche dieser Regeln in die Ergebnismenge kommen. Um nicht alle Kombinationen berechnen zu müssen, gibt es z.B. den Apriori-Algorithm.

Ferienclub

Als mögliche Anwendungsfälle für Business Intelligence im Projekt „Ferienclub“ wäre das Erstellen einer personenbezogenen Eventliste. Anstelle dass der Kunde sich durch die Vielzahl an Angeboten durcharbeiten muss, kann hier anhand der bekannten Informationen über ihn (Herkunft, Alter, Begleiter, etc.) und das Verhalten anderer Besucher eine Liste mit ausgewählten Angeboten bereitgestellt werden. Z.B. könnten bei Angeboten für Eltern gleich mit angegeben werden, wo ihre Kinder in der Zwischenzeit beschäftigt werden können.

Eine weitere Möglichkeit wäre es die Animatoure über Kunden zu informieren, ob diese zu bestimmten Zeiten ihre Ruhe haben oder beschäftigt werden möchten, indem Daten herangezogen werden, wie der Kunde auf den Animateur reagiert hat.

Auch können Auswertungen gemacht werden, um zu sehen wo sich verschiedene Gruppen zur gleichen Zeit aufhalten und zur selben Zeit andere Bereiche kaum genutzt werden, aber auch, ob es bestimmte Zeiten gibt, an denen es für bestimmte Kunden keine interessanten Angebote gibt. Diese Informationen sind u.a. nützlich, um die Auslastung der Bereiche gleichmäßig zu verteilen.

Meine Motivation

Wie oben bereits erwähnt wird mein Beitrag für das Projekt sein, mich um den Analyseteil beim Business Intelligence zu kümmern. Ich habe mich hierfür entschieden, weil ich das Thema Business Intelligence interessant finde und vor allem was sich aus den vorhandenen Daten an neuen Informationen herleiten lässt.

Ein weiterer Grund für Business Intelligence ist ein Projekt an dem ich beteiligt war. Ich habe dort bereits unbewusst einen Teil dieser Techniken angewandt.

Andere Projektteile

Am engsten werde ich mit Jan Weinschenker zusammenarbeiten, da er sich um den Datenimport kümmert. Des Weiteren werde ich mit den Kommilitonen, die sich mit Service Oriented Architecture beschäftigen, zusammenarbeiten, um meinen Dienst anbieten zu können. Weitere Projektteile wären Persistenz und RFID, um Daten bereitgestellt zu bekommen. Diese werden auf [ANW1] vorgestellt.

Literaturverzeichnis

- [ANW 1] Vorträge Master Informatik HAW Hamburg 2005 (Anwendungen 1)
Im Internet zu finden unter:
<http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2005/vortraege.html> (6.Juli 2005)
- [BG 04] Andreas Bauer, Holger Günzel (Hrsg.):
Data-Warehouse-Systeme, 2. Aufl. Dpunkt, 2004 –
ISBN 3-89864-251-8
- [MMC 05] John Medicke, Margie Mago, Feng-Wei Chen:
Creating an Intelligent and Flexible Solution with BPM, Business Rules,
and Business Intelligence: Part 3: Data Warehouse Visibility,
20.November 2003. Im Internet zu finden unter:
<http://www-128.ibm.com/developerworks/db2/library/techarticle/dm-0311chen/> (14.Mai 2005)
- [WIKI 05] Wikipedia: Business Intelligence, 17.Mai 2005.
Im Internet zu finden unter:
http://de.wikipedia.org/wiki/Business_Intelligence (17.Mai 2005)