Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Ausarbeitung Projekt

Sven Elvers

Business Intelligence: Analyse

Betreuender Prüfer: Prof. Dr. Olaf Zukunft

Fakultät Technik und Informatik
Studiendepartment Informatik

Sven Elvers

Thema der Ausarbeitung
Projekt

Business Intelligence: Analyse

Stichworte

Business Intelligence, Analyse, Data Warehouse, Report, OLAP, Pentaho, BIRT, Mondrian, OpenI

Kurzzusammenfassung

In diesem Artikel wird über die Erfahrungen und die Ziele berichtet, die im Projekt "Ferienclub" bei der Analyse in der Teilgruppe "Business Intelligence" gemacht bzw. erreicht wurden.

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation

In der Industrie sowie der Wissenschaft werden heutzutage viele Informationen gespeichert, wie zum Beispiel über Kundenverhalten oder Messergebnisse. Diese Daten werden dann analysiert, um z.B. den Gewinn oder die Auslastung der Ressourcen zu ermitteln, und graphisch dargestellt. Zusätzlich können in diesen Daten Informationen enthalten sein, die nicht offensichtlich sind, aber zu neuen Erkenntnissen führen oder zur Verbesserung von Strategien genutzt werden können.

Diese Informationen lassen sich mit Techniken und Technologien aufdecken, die unter dem Begriff *Data Mining* zusammengefasst sind. Es werden unter anderem statistische Verfahren angewendet oder intelligente Systeme eingesetzt.

Um Berichte zu erstellen, die Daten dynamischen zu analysieren oder neue Informationen zu gewinnen, werden Business Intelligence Systeme eingesetzt. Da diese die Daten aus den Produktionssystemen extrahieren, bereinigen (Fehlerkorrektur, doppelte Einträge entfernen, etc.) und in einer für die Analyse effizienter Form speichern. Auf diesen Daten können dann die Auswertungen getrennt von den Produktionssystemen laufen, um diese nicht zu beeinträchtigen bzw. selbst beeinträchtigt zu werden.(vgl. Bauer (2004), Otte und Nathansen)

1.2 Ziele

Das Ziel in dem Projekt war zusammen mit Jan Weinschenker (Wintersemester 2005/06) ein Business Intelligence System aufzusetzen. Dieses System sollte Daten, die im Ferienclub anfallen, sammeln, aufbereiten und analysieren. Meine Aufgabe dieses Teilprojektes war, den Analyseteil des Business Intelligence Systems zu realisieren. Die Basisdatenbank wurde als Schnittstelle zwischen den beiden Teilbereichen (siehe Abschnitt 2) definiert. Um das gemeinsame Ziel zu erreichten, haben wir folgende Meilensteine definiert.

Meilenstein I Als ersten Meilenstein sollte die Arbeitsumgebung aufgesetzt werden. Die Werkzeuge für die Entwicklung und die Komponenten für das Business Intelligence System, sollten bis zu diesem Punkt laufen, um früh sicher zu sein, das diese funktionieren bzw. die Komposition dieser Komponenten.

Meilenstein II Das Ziel des zweiten Meilensteins war, die gewählten Komponenten zu testen, ob diese geeignet sind. Dazu sollten Datenmodelle mit kleinen Unterschieden erstellt und mit Testdaten gefüllt werden. Anhand des Integrieren dieser Daten und das Erstellen erster Analysen sollte sichergestellt werden, das die Realisierung des Business Intelligence Systems mit diesen Komponenten erreicht werden kann.

Meilenstein III Beim dritten Meilenstein sollten dann Daten aus einem im Ferienclub vorhandenen System extrahiert, transformiert und dann in die Basisdatenbank geladen werden. Auf diesen Daten sollten dann erste Analysen laufen und Reports erstellt werden. Die Daten in der Basisdatenbank sollen in einem Starschema strukturiert sein (siehe Abschnitt 3).

2 Business Intelligence

Unter dem Begriff Business Intelligence wird im allgemeinen die analytischen Konzepte, Prozesse und Werkzeuge zusammengefasst. Ziel ist es, aus vorhandenen Daten neues Wissen zu erlangen. Dieses wird dann u.a. im betriebswirtschaftlichen Bereich für Entscheidungen herangezogen oder im wissenschaftlichen Bereich, um empirisch ermittelte Daten zu interpretieren.

Business Intelligence Systeme sind wie in Abbildung 1 aufgebaut. Die Daten werden aus den Systemen extrahiert (Datenbeschaffungsbereich) und in einer separaten Datenbank gespeichert, auf denen dann die Analysen laufen.(vgl. Bauer (2004), Wikipedia BI 2005)

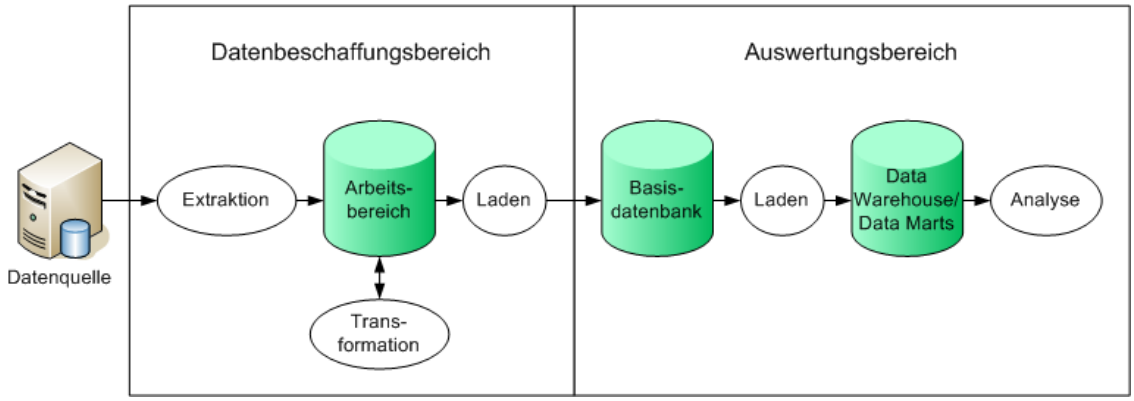


Abbildung 1: BI-Architektur

2.1 Datenquellen

Das Business Intelligence System holt sich die Daten aus unternehmensinternen oder -externen Datenquellen. Diese können z.B. relationale Datenbanken oder Log-Dateien sein. Die Datenquellen sind aber kein Bestandteil des Business Intelligence Systems.

2.2 Arbeitsbereich

Bevor die Daten für die Analysen in die Basisdatenbank kommen, werden sie zur Überarbeitung im Arbeitsbereich gespeichert. Für diese Vorgänge sind die ETL-Komponenten verantwortlich (**E**xtraktion, **T**ransformation, **L**aden).

Da die Informationen in den Datenquellen unterschiedlich gespeichert oder Fehlerhaft sein können, muss die Transformation Datentypen anpassen, Kodierungen konvertieren, Zeichenketten und Datumsangaben vereinheitlichen, Maßeinheiten umrechnen und Attributwerte kombinieren oder separieren, wie z.B. Name und Nachname von Personen.

Für die Fehlerbehandlung können verschiedene Strategien genutzt werden. Diese sind nötig bei fehlerhaften, redundanten, veralteten oder fehlenden Werten nötig. Eine Strategie könnte zum Beispiel sein, das die Informationen aus einer bestimmten Datenquelle bevorzugt genutzt werden soll.

2.3 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-Speiche-Architektur), wodurch die Anzahl der Schnittstellen von m*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.

2.4 Data Warehouse/ Data Marts

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 loaderßelbst. Der "bulk loaderïst 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.

2.5 Analyse

Bei den Analyseverfahren werden drei Hauptkategorien unterschieden:

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

Data Access Diese Art von Analysen werden von Reporting Tools verwendet. Diese beschränken sich auf das Lesen der Daten und der Repräsentation dieser in einem Bericht. Eine häufig genutzter Dienst dieser Tools ist die *Ampelfunktion*. Hier werden Grenzwerte und Darstellungen angegeben, um Besonderheiten hervor zu heben. Eine klassische Variante ist, Werte in grün, gelb oder rot anzuzeigen.

OLAP Bei Tools, die diese Analyse unterstützen, braucht der Benutzer nur einmal am Anfang alle Kriterien zu definieren und kann sich dann in dem Dimensionswürfel des Data Marts bewegen (siehe Abb. 2). 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).

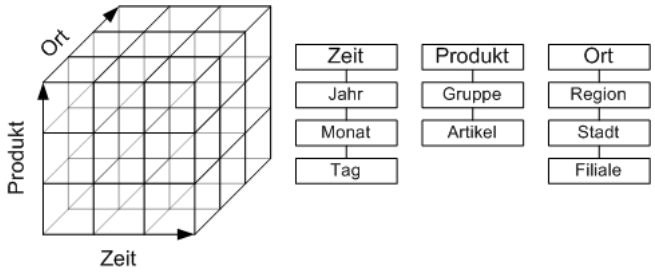


Abbildung 2: Dimensionswürfel

Data Mining 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 vorhandenen 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.

3 Basisdatenbank

Bei der Datenbank haben wir uns für BizGres (BizGres 2005) mit dem Betriebssystem Fedora (Fedora 2005) entschieden. Dieses DBMS baut auf PostgreSQL auf und ist speziell für Business Intelligence Systeme, um diese gezielt zu unterstützen. Ein weiterer Grund für BizGres war, dass eine ausführliche Dokumentation vorhanden ist.

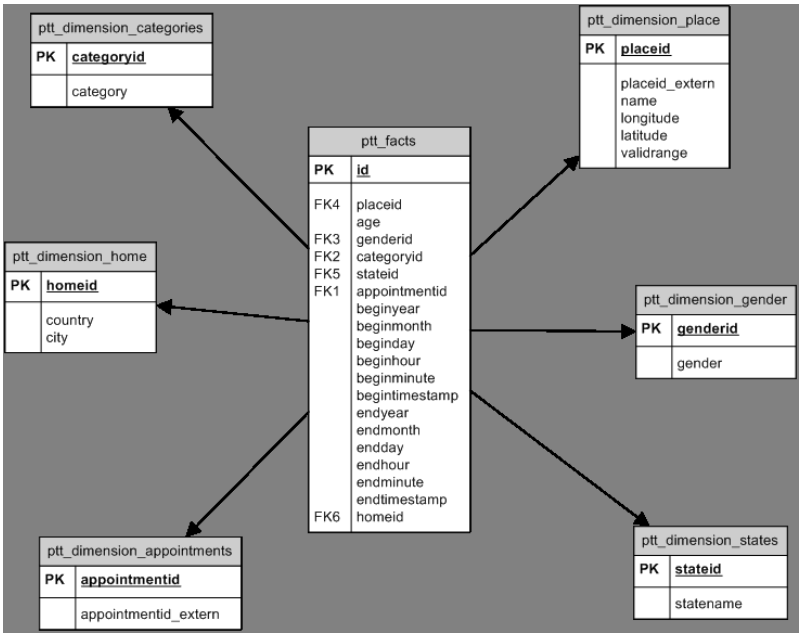


Abbildung 3: Starschema

Anstelle einer Basisdatenbank und eines Data Warehouse zu nutzen (siehe Abschnitt 2), haben wir uns dafür entschieden, nur eine Datenbank zu realisieren, die als Basis für die

Analysen genutzt werden sollte. Der Schritt, die Daten aus der Basisdatenbank zu extrahieren und in grobgranularer oder vorgefilterter Form in ein Data Warehouse zu schreiben, war für unsere Analysen nicht erforderlich, da die Daten, wie sie in der Basisdatenbank vorhanden sind, erforderlich waren.

In Abbildung 3 auf Seite 5 ist das Starschema zu sehen, wie die Daten in der Basisdatenbank strukturiert sind. Dieses Schema wird von Jan Weinschenker (Wintersemester 2005/06) mit den Daten aus der Datenbank des Terminplaners von Mark Thomé (Wintersemester 2005/06) gefüllt.

Im Zentrum ist die Faktentabelle zu sehen. In dieser werden die Informationen zu einem Appointment gespeichert. Als Dimensionen werden in separaten Tabellen die Kategorien (ptt_dimension_categories), der Veranstaltungsort (ptt_dimension_place), das Geschlecht des Benutzers (ptt_dimension_gender), der Status des Appointment (ptt_dimension_states) und der Heimatort des Benutzers (ptt_dimension_home) gespeichert. Die Dimension Zeit ist direkt in der Faktentabelle, da diese selbst eindeutig ist. Die Tabelle ptt_dimension_appointments wird nicht für die Analysen benötigt, sondern ist für die Datenbeschaffung da, um die Datensätze mit den jeweiligen in den Datenquellen zuordnen zu können.

4 Reporting

Obwohl Pentaho (Pentaho 2005) nur als Pre-Release Version zur Verfügung stand, habe ich mich dafür entschieden, diese zu nutzen. Denn es gab einen detaillierten Meilensteinplan, anhand dessen ich sehen konnte was implementiert ist und wie zuverlässig sie diesen verfolgten. Des Weiteren gab es ausführliche Dokumentationen.

Um Pentaho mit neuen Berichten zu erweitern, musste ich für jeden neuen Bericht zwei Dateien erstellen. Mit der ersten Datei (rptdesign-Datei) wurde beschrieben, wie der Bericht aufgebaut sein soll, welche Daten er aus welcher Datenbank holen muss, welche er von der Anwendung oder dem Benutzer bekommt und wie die Daten verarbeitet werde müssen. Die zweite Datei (xaction-Datei) brauchte Pentaho, um zu wissen, wie die rptdesign-Datei heißt, welche Informationen er für den Bericht braucht, für eine allgemeine Beschreibung des Berichts und weiteren Parametern die für die Erstellung des Berichts benötigt wurde.

Beide Dateien konnte ich mit Eclipse[1] Plug-In's erstellen. Für die xaction-Datei ist das Plug-In bei Pentaho dabei und es müssen nur Felder ausgefüllt werden. Um die rptdesign-Datei zu erstellen, habe ich das Plug-In BIRT (BIRT 2005) verwendet. Dieses bietet eine graphische Oberfläche, um das Layout für den Bericht zu erstellen und um die Datenquellen, die Datensätze und Berichtsparameter anzugeben (siehe Abb. 4 und 5).

Für Pentaho habe ich zwei Berichte erstellt. Beim ersten Bericht (Abb. 6) wird die Auslastung der Orte im Ferienclub präsentiert bzw. wie viele Appointments den jeweiligen Ort als

[1] `http://www.eclipse.org` (14.Februar 2005)

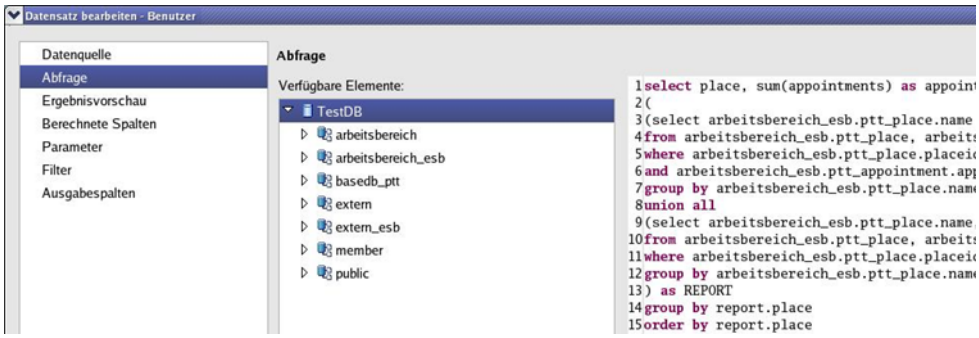
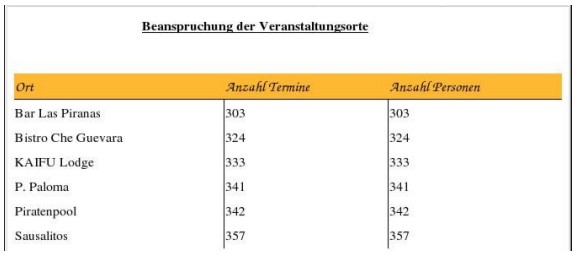


Abbildung 4: BIRT - Datensatz

Abbildung 5: BIRT - Layout

Treffpunkt haben. Beim zweiten Bericht (Abb. 7) muss der Anwender als Parameter ein Ort auswählen, zu dem die Altersverteilung ausgegeben werden soll. Zusätzlich zu der Tabelle sollte hier auch eine Grafik ausgegeben werden. Dieses habe ich aber nicht geschafft zu realisieren, da Fehlermeldungen weder ausgegeben wurden noch im Log waren.



Beanspruchung der Veranstaltungsorte

Ort	Anzahl Termine	Anzahl Personen
Bar Las Piranas	303	303
Bistro Che Guevara	324	324
KAIFU Lodge	333	333
P. Paloma	341	341
Piratenpool	342	342
Sausalitos	357	357

Abbildung 6: Auslastung der Veranstaltungsorte

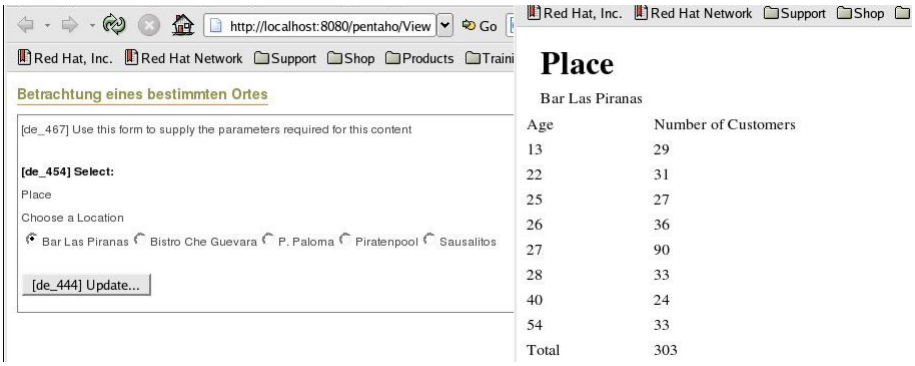


Abbildung 7: Altersverteilung eines Ortes

5 Online Analytical Processing (OLAP)

Bei der Auswahl eines geeigneten OLAP-Tools hatte ich mich für die Kombination von Mondrian (Mondrian 2005) und OpenI (OpenI 2005) entschieden. Mit Mondrian wird ein OLAP Server aufgesetzt, der die Funktionen bereitstellt und auf die BizGres-Datenbank zugreifen kann. Um dem Anwender ein Interface zur Verfügung zu stellen, wird OpenI verwendet. Diese Webanwendung bindet Mondrian ein und bietet die erforderliche Navigation an. Des Weiteren sind beide Anwendungen ausführlich dokumentiert.

```
<Dimension name="End Time">
        <Hierarchy hasAll="true" primaryKey="id">
                <Table schema="basedb_ptt" name="ptt_facts"/>
                <Level name="Year" column="endyear" type="Numeric" uniqueMembers="true"/>
                <Level name="Month" column="endmonth" uniqueMembers="false" type="Numeric"/>
                <Level name="Day" column="endday" uniqueMembers="false" type="Numeric"/>
                <Level name="Hour" column="endhour" uniqueMembers="false" type="Numeric"/>
                <Level name="Minute" column="endminute" uniqueMembers="false" type="Numeric"/>
        </Hierarchy>
</Dimension>
<Cube name="Appointments">
        <Table schema="basedb_ptt" name="ptt_facts"/>
        <DimensionUsage name="Guest Home" source="Guest Home" foreignKey="homeid"/>
        <DimensionUsage name="Appointment States" source="Appointment States" foreignKey="stateid"/>
        <DimensionUsage name="Appointment Places" source="Appointment Places" foreignKey="placeid"/>
        <DimensionUsage name="Appointment Categories" source="Appointment Categories" foreignKey="categoryid"/>
        <DimensionUsage name="Appointment IDs" source="Appointment IDs" foreignKey="appointmentid"/>
        <DimensionUsage name="Gender" source="Gender" foreignKey="genderid"/>
        <DimensionUsage name="Begin Time" source="Begin Time" foreignKey="id"/>
        <DimensionUsage name="End Time" source="End Time" foreignKey="id"/>

        <Measure name="Appointments" column="id" aggregator="count" formatString="#,###"/>
        <Measure name="Youngest Guest" column="age" aggregator="min" formatString="#,###"/>
        <Measure name="Oldest Guest" column="age" aggregator="max" formatString="#,###"/>
</Cube>
```

Abbildung 8: Mondrian - Beschreibung des Starschemas

Beim Mondrian OLAP Server musste die Datenbankverbindung eingerichtet und eine

XML-Datei erstellt werden, die das Datenbankschema (Starschema) beschreibt. In Abbildung 8 auf Seite 8 ist ein Auszug zu sehen, wie das Starschema beschrieben ist. Bei den Dimensionen wird die Tabelle mit dem Primary Key angegeben und vom grobgranularem Level zum feingranularen Level die Attribute. Auch beim Cube wird die Tabelle und der Primary Key angegeben. Aber statt den Leveln werden hier die einzelnen Dimensionen zusammen mit dem zugehörigen Foreign Key sowie den Metriken angegeben. Die Dimensionen können auch direkt in der Definition für den Cube angegeben werden, sind dann aber nicht von weiteren Cubes verwendbar.

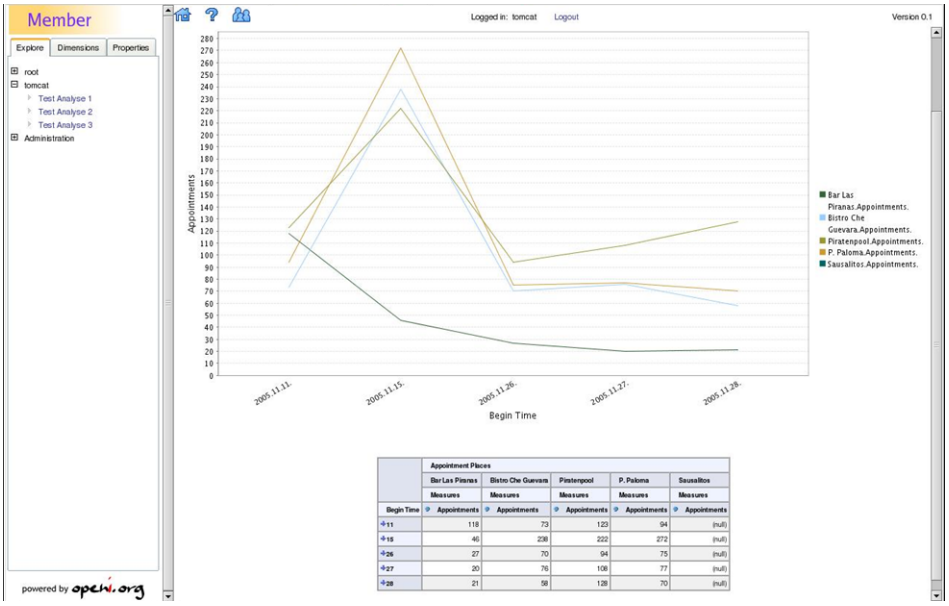


Abbildung 9: OpenI - Zeitliche Betrachtung der Appointments zu den Orten

Für OpenI musste ein neues Projekt erstellt werden, welches den neu definierten Catalog von Mondrian benutzt. Neue Benutzer und Rollen wurden eingerichtet. Des Weiteren habe ich drei Sichten definiert, mit denen der Benutzer anfangen kann:

- In Abhängigkeit zu den Orten werden in einem Säulendiagramm die Anzahl der Appointments, der älteste und der jüngste Besucher ausgegeben.
- In Abhängigkeit zu den Orten und dem Appointmentstatus werden in einem Säulendiagramm die Anzahl der Appointments angezeigt.
- In zeitlicher Abhängigkeit werden zu den Orten die Anzahl der Appointments in einem Liniendiagramm angezeigt (siehe Abb. 9).

Die Navigation erfolgt in der Tabelle. Hier kann nach Spalten sortiert werden oder je nach vom Benutzer erfolgte Einstellung ein Drill-down oder Drill-up durchgeführt werden, indem auf den Zeilennamen geklickt wird. Weitere Möglichkeiten sind u.a. den Bericht als PDF-Dokument zu speichern, die Dimensionen tauschen oder einen anderen Graphen auszuwählen.

6 Zusammenfassung und Ausblick

Den ersten Meilenstein habe ich eingehalten. Die Tools mit den Beispielen liefen zu dem festgelegten Termin. Den zweiten Meilenstein habe ich erst eine Woche später erreicht, zu realisieren. Diese Zeit konnte ich bis zum dritten Meileinstein wieder aufholen. Das Ziel den Analysebereich des Business Intelligence Systems zu realisieren, habe ich erreicht. Es können für die Daten von der Anwendung von Mark Thomé (Wintersemester 2005/06) Berichte erstellt werden und mit dem OLAP-Tool OpenI analysiert werden.

Die Trennung zwischen Anwendung, Datenmodell und Analyse ist sowohl bei Pentaho als auch bei OpenI gut durchdacht, so dass keine Anwendungsprogrammierung nötig war. Man konnte sich auf die Abläufe in dem Business Intelligence System konzentriert werden, welche Daten stehen zur Verfügung, wie sollte das Starschema aussehen, welche Analysen sollen auf den Daten laufen und wie sollen die Ergebnisse dargestellt werden. Die meiste Zeit nahm das aufsetzen der Server in Anspruch, sobald die ersten Analysen auf den eigenen Daten liefen, ließen sich beide Tools einfach erweitern. Zeitaufwändig war auch die Fehlersuche, da die Fehlermeldungen meist nicht hilfreich waren.

Da Pentaho während der Laufzeit des Projekts weiterentwickelt wurde, wäre es interessant zu untersuchen, was es kann und wo seine Grenzen liegen. Eine Möglichkeit ist den OLAP-Teil mit Pentaho zu realisieren, da Mondrian jetzt eine seiner Komponenten ist. Es wäre auch interessant diese Anwendung um Data Mining Analysen oder die Rechteverwaltung des OLAP-Servers zu erweitern, dass zum Beispiel bestimmte Personen nur einen Aspekt einer Dimension sehen dürfen.

Literatur

[Bauer 2004] BAUER, Andreas ; GÜNZEL, Holger (Hrsg.): *Data Warehouse Systeme, 2.Aufl.* dpunkt, 2004. – ISBN 3-89864-251-8

[BIRT 2005] *BIRT, Business Intelligence and Reporting Tools*. – URL `http://www.eclipse.org/birt/`. – (14. Februar 2005)

[BizGres 2005] *BizGres, PostgreSQL for Business Intelligence and Data Warehousing*. – URL `http://www.bizgres.org`. – (14. Februar 2005)

[Fedora 2005] *Fedora Linux*. – URL `http://fedora.redhat.com`. – (14. Februar 2005)

[Mondrian 2005] *Mondrian OLAP Server*. – URL `http://mondrian.sourceforge.net`. – (15. Februar 2005)

[OpenI 2005] *OpenI (Open Intelligence), Open Source Web Application for OLAP Reporting*. – URL `http://openi.sourceforge.net`. – (15. Februar 2005)

[Otte und Nathansen] OTTE, Ralf ; NATHANSEN, Martin: *Data Mining Applications for Industry Germany*. – URL `http://www.datmin.de/index.html`. – (28. Dezember 2005)

[Pentaho 2005] *Pentaho, open source business intelligence*. – URL `http://www.pentaho.org`. – (14. Februar 2005)

[Thomé Wintersemester 2005/06] THOMÉ, Mark: Projekt Ferienclub — Pocket Task Timer - A Personal Approach on Location-Based Services / HAW Hamburg — Masterprogramm Verteilte Systeme. Wintersemester 2005/06. – Forschungsbericht

[Weinschenker Wintersemester 2005/06] WEINSCHENKER, Jan: Projekt Ferienclub — Business Intelligence: Extraktion, Transformation, Laden / HAW Hamburg — Masterprogramm Verteilte Systeme. Wintersemester 2005/06. – Forschungsbericht

[Wikipedia BI 2005] *Wikipedia: Business Intelligence*. – URL `http://de.wikipedia.org/wiki/Business_Intelligence`. – (17.Mai 2005)