ETL-Tool Survey Light: Ein Vergleich des OWB mit Open Source ETL-Tools

Prof. Dr. Reinhold von Schwerin

Hochschule Ulm, Fakultät für Informatik

15.09.2011

Schlüsselworte: BI in der Lehre, ETL-Prozess, ETL-Tool
Schlüsselworte: Oracle Warehouse Builder, Talend Open Studio, Pentaho Data Integration, Clover ETL

Zusammenfassung

Im Rahmen einer Lehrveranstaltung wurde von Studierenden der Wirtschaftsinformatik an der Hochschule Ulm im Sommersemseter 2011 ein Vergleich von ETL-Tools durchgeführt. Jede Studierendengruppe verglich den Oracle Warehouse Builder mit einem Open Source Tool auf Basis eines Teils der Kriterien des ETL-Tool Survey. Anhand einer realitätsnahen Aufgabenstellung zeigte sich, dass die Werkzeuge Talend Open Studio und Pentaho Data Integration gut mit dem OWB mithalten konnten. Dahingegen konnte das Tool Clover ETL hauptsächlich bezüglich der geringen Einstiegshürden punkten.

Einleitung

Im Studiengang Wirtschaftsinformatik an der Hochschule Ulm steht im 5. Semester die Veranstaltung *Data Warehousing* auf dem Lehrplan. Die Lehre erfolgt anwendungsorientiert anhand einer Fallstudie, die vom *DELL DVDStore* (siehe [1]) abgeleitet wurde. Die inhaltlichen Schwerpunkte liegen dabei auf *multidimensionalen Konzepten* (insbes. Modellierung), Auswertungen mit *SQL/OLAP* (siehe auch [4]) und dem *ETL-Prozess* (inkl. Stored Procedures).
Die ETL-Aufgabe besteht aus der Integration der operativen MySQL-Datenbanken zweier DVD-Online Shops und zusätzlicher Informationen der (fiktiven) Website *The Movie Site* gemäß Abbildung 1. Diese Aufgabe enthält typische ETL-Teilaufgaben wie das *Filtern* der relevanten Daten sowie insbesondere Aufgaben der *Harmonisierung* wie den Einsatz von Zuordnungs- und Nachschlagetabellen, die Auflösung von Schlüsseldisharmonien sowie den Umgang mit Synonymen und Kodierungsunterschieden. Darüberhinaus sind auch gewisse *Anreicherungen* nötig (siehe etwa [3] für die zugehörige Theorie).

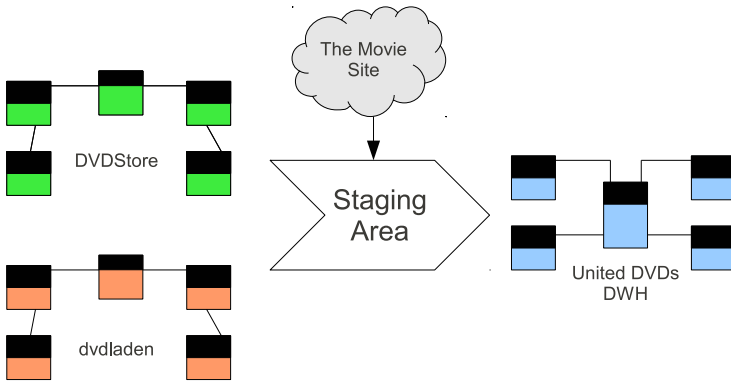

Abbildung 1: ETL-Aufgabe der Data Warehousing Fallstudie

Gelöst wird die Aufgabe zunächst mit Hilfe selbst erstellter SQL-Skripte mit Stored Procedures, wobei die Zieldatenbank zunächst wieder eine MySQL-Datenbank in Form eines Sternschemas ist. Diese wird anschließend in eine Oracle-Datenbank migriert, da MySQL nur `ROLLUP` als SQL/OLAP-Erweiterung bietet, Oracle hingegen eine ganze Reihe weiterer `GROUP BY` Erweiterungen, etwa `CUBE`, sowie *analytische Funktionen*, insbesondere *Windowing* und *Ranking*.
Die Studierenden müssen im Fach Data Warehousing weiterhin eine Prüfungsvorleistung erbringen. Diese besteht immer aus der Erarbeitung eines vorgegebenen Themas sowie der Abgabe eines Dokuments, welches eine Teilnehmergruppe von 4-5 Personen gemäß den formalen Anforderungen an eine *wissenschaftliche Arbeit* erstellt. Im Sommersemester 2011 bestand die Aufgabe darin, den vorgenannten ETL-Prozess nochmals unter Verwendung zweier ETL-Tools durchzuführen und diese gleichzeitig anhand vorgegebener Kriterien zu evaluieren. Eines der Werkzeuge war jeweils ein Open Source Tool und das andere der Oracle Warehouse Builder (OWB). Die Nutzung der Oracle Datenbank und der zugehörigen Tools wie dem OWB erfolgt dabei mit einer Lizenz der Oracle Academy. Die Nutzung der Open Source Tools ist frei, wobei es neben der hier verwendeten Community Edition jeweils auch eine kostenpflichtige Enterprise Edition mit Support für die Tools gibt. Im einzelnen handelt es sich um

- Pentaho Data Integration (PDI)
- Talend Open Studio (TOS)
- Clover ETL

All diese Tools werden auch im aktuellen ETL-Tool Survey auf www.etltool.com bewertet, wobei letzteres insgesamt 19 ETL-Tools anhand einer fixen Kriterienliste vergleicht. Für die studentische Aufgabe wurde lediglich ein Ausschnitt dieser Kriterien verwendet. Diese werden im folgenden Abschnitt näher erläutert. Abschließend werden die Vergleichsergebnisse der Studierenden [2] zusammenfassend dargestellt und ein Fazit gezogen.

Vergleichskriterien und Gewichtung

Gegenüber den Kriterien des ETL-Tool Surveys wurden die Aspekte *Architektur* und *Echtzeitfähigkeit* ausgeklammert, da diese mit der oben dargestellten Aufgabenstellung und der vorhandenen Infrastruktur nicht wirklich untersucht werden konnten. Ebenso sollten Installationsaspekte

keine Rolle spielen, obwohl die Studierenden die Open Source Tools selbst installieren mussten, während Ihnen der OWB und die Oracle Datenbank vorkonfiguriert in einer virtuellen Maschine zur Verfügung standen.
Die tatsächlich bewerteten Kriterien inklusive der zugehörigen Unterkriterien werden im Folgenden erläutert. Dabei ergibt sich die Bewertung der Kriterien als Mittelwert der Bewertung der Unterkriterien, wobei nicht notwendiger Weise jedes Unterkriterium eine Bewertung erhalten muss. Pro Unterkriterium können maximal 10 Punkte vergeben werden. Die Gewichte bewegen sich zwischen 1 und 5.
Im Einzelnen werden *ETL-Funktionalität, Benutzerfreundlichkeit, Wiederverwendbarkeit, Debugging* und *Konnektivität* mit den im Folgenden erläuterten Unterkriterien betrachtet. Dabei werden lediglich die zu denjenigen Unterkriterien gehörenden Fragen näher ausgeführt, die nicht selbsterklärend sind.

ETL-Funktionalität

- *Aufteilen von Datenströmen/Verschiedene Ziele*
  Besteht die Möglichkeit eine Datenquelle einmal einzulesen und die Ergebnisse dann auf mehrere Tabellen aufzuteilen?
- *Bedingtes Aufteilen*
  Kann die genante Aufteilung auch gemäß vordefinierter Bedingungen erfolgen (z.B. „Stammt der Kunde aus Region A, dann schreibe die Daten in Tabelle 1, sonst in Tabelle 2“)?
- *Vereinigung*
  Können gleichstrukturierte Datensätze aus verschiedenen Quellen in die gleiche Tabelle geschrieben werden?
- *Schlüsselsuchfunktion im Speicher*
  Können Tabellen komplett in den Hauptspeicher geladen und dann dort gesucht werden?
- *Wiederverwendbarkeit der Schlüsselsuchfunktionen für mehrere Arbeitschritte*
- *Slowly Changing Dimensions*
- *Scheduler*
  Gibt es einen Scheduler, der auch Abhängigkeiten berücksichtigt?
- *Job – Fehlerbehandlung*
  Kann innerhalb eines Jobs auf Fehler reagiert werden?
- *Auswirkungsanalyse*
  Kann man die Auswirkung geplanter Änderungen a-priori untersuchen?
- *Datenabstammung*
  Lässt sich die Quelle eines Informationselements bestimmen (inverse Auswirkungsanalyse)?
- *Automatische Dokumentation*
- *Unterstützung von Data Mining Modellen*

Da für ein ETL-Tool natürlich die gebotene Funktionalität das wichtigste Kriterium darstellt wird dieses Kriterium mit 5 gewichtet.

Benutzerfreundlichkeit

- *Bedienbarkeit*
  Ist das Tool leicht zu erlernen und täglich zu benutzen?
- *WYSIWYG*
  Wird das WYSIWYG-Prinzip auf die Daten angewandt?
- *GUI-Design*
- *Lernaufwand*

Da dieses Kriterium natürlich insbesondere für Einsteiger wie die Studierenden wichtig ist, erhält es das Gewicht 4.

Wiederverwendbarkeit

- *Wiederverwendbarkeit von Komponenten*
- *Dekomposition*
  Lässt sich der ETL-Workflow modular zusammensetzen?
- *Benutzerdefinierte Funktionen*
  Kann selbst geschriebener Code in den Workflow mit eingebunden werden?
- *Bemerkungen bei Datenobjekten*
  Können Datenobjekte gemeinsam kommentiert werden, so dass diese fest verbunden bleiben?

Dieses Kriterium wurde als nicht ganz so wichtig erachtet und daher nur mit dem Faktor 2 versehen.

Debugging

- *Schritt für Schritt Debugging*
- *Zeile für Zeile Debugging*
- *Breakpoints*
  Lassen sich für bestimmte Prozessschritte bzw. Datenreihen Breakpoints setzen?
- *Watchpoints*
  Lässt sich der Ablauf bei Eintreten bestimmter Bedingungen unterbrechen?
- *Compiler/Validator*
  Kann man den Prozess bzw. Code einfach validieren und werden Fehler gemeldet und markiert?

Dieses Kriterium wird für einigermaßen wichtig erachtet und daher mit dem Faktor 3 gewichtet.

Konnektivität

- *Native Konnektivität*
  Wieviele und welche nativen Verbindungen bietet das Tool (ausgenommen ODBC, OLE DB und Textdateien)?
- *Unterstützung für Joined Tables als Quelle*
  Lassen sich Tabellen graphisch verbinden und der Join wird dann auf der Datenbank anstelle vom ETL-Tool ausgeführt?
- *Veränderungen in den Quelldaten*
  Kann man sicherstellen, dass nur die in der Quelle geänderten Daten übertragen werden (*changed data capture*)?
- *Verfügbarkeit von Funktionen zur Sicherung der Datenqualität*
- *Verfügbarkeit von Funktionen zur Datenvalidierung*
- *Datenbeschreibung*
  Lassen sich typische Aufgaben des *Data Profiling* mit dem Tool erledigen?

Da dieses Kriterium maßgeblich die Einsatzmöglichkeiten des Tools bestimmt, erhält es das Gewicht 4. Somit sind nun die Kriterien näher ausgeführt und der folgende Abschnitt beschäftigt sich mit den Ergebnissen der Evaluierung.

Ergebnisse des Toolvergleichs

In Tabelle 1 sind die ungewichteten Ergebnisse als *Prozentwerte* zusammen mit den Gewichten dargestellt. Die Zahlen ergeben sich als Mittelwert über die Bewertungen der jeweiligen Kriterien der einzelnen Gruppen, wobei sich die Bewertung der Kriterien wiederum als das Mittel der Bewertung der Unterkriterien ergibt. Konkret heißt dies, dass die Werte des OWB über 5 Gruppen, die von TOS und PDI über je 2 Gruppen gemittelt sind, während Clover ETL nur von einer Gruppe verwendet wurde, deren Bewertung direkt wiedergegeben wird.

Kriterium \Tool	*Gewicht*	PDI	TOS	Clover	OWB
ETL-Funktionalität	*5*	69	72	22	80
Benutzerfreundlichkeit	*4*	71	52	80	57
Wiederverwendbarkeit	*2*	64	64	35	72
Debugging	*3*	89	76	24	90
Konnektivität	*4*	84	79	17	89

Tabelle 1: Ergebnisübersicht des ETL-Toolvergleichs

Abbildung 2 verdeutlicht die ungewichteten Ergebnisse des Vergleichs noch einmal visuell. Es ist deutlich zu erkennen, dass OWB und PDI die Nase vorne haben, während Clover ETL nicht wirklich eine Konkurrenz darstellt.

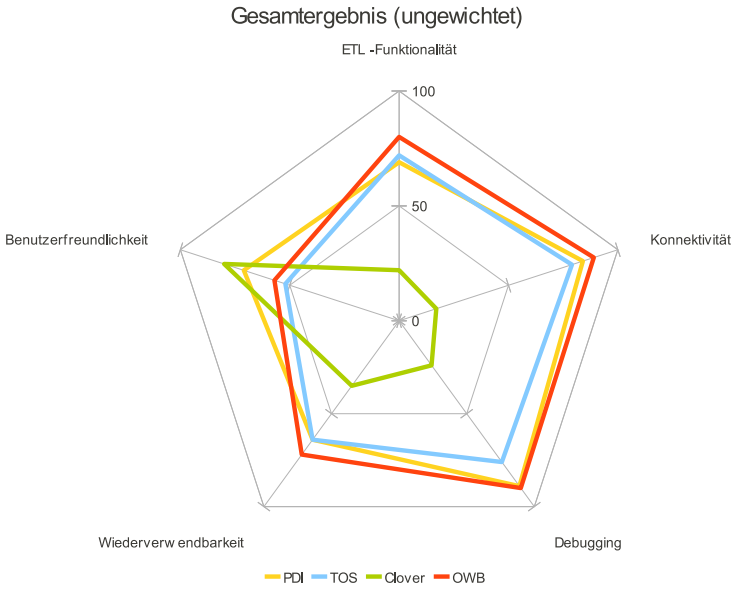

Abbildung 2: Vergleichsergebnisse (ungewichtet)

Unter Berücksichtigung der Gewichtung ergibt sich das Diagramm in Abbildung 3. Auch hier zeigt sich sehr deutlich, dass Clover ETL beinahe die Hälfte seiner Gesamtpunkte der leichten Bedienbarkeit verdankt, insgesamt aber gegenüber den anderen untersuchten Tools klar zurückbleibt. Der OWB geht im Vergleich zu PDI und TOS (knapp) als Sieger hervor, was insbesondere auf die in den Augen der Studierenden größere vorhandene ETL-Funktionalität zurückzuführen ist, während PDI unter anderem auch als ähnlich leicht bedienbar erlebt wird wie Clover ETL und dadurch im Gesamtergebnis kaum hinter dem OWB liegt.

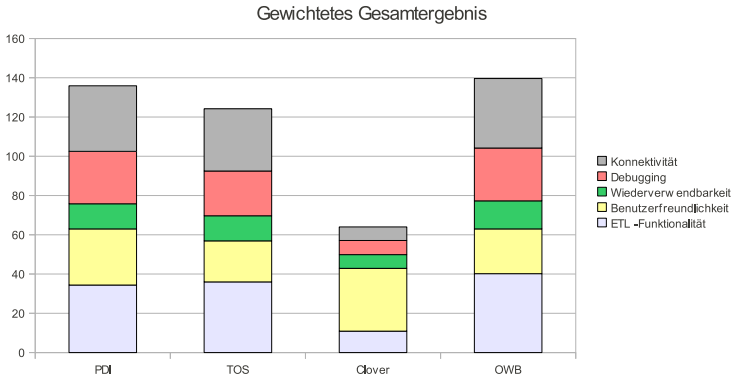

Abbildung 3: Vergleichsergebnisse (gewichtet)

Fazit

Im Sommersemester 2011 konnten die Studierenden des 5. Semesters des Studiengangs Wirtschaftsinformatik an der Hochschule Ulm anhand eines realistischen Szenarios ein Open Source ETL-Tool mit dem OWB vergleichen. Neben dem Ausbildungsgehalt des praktischen Umgangs mit ETL-Tools enstand somit eine Evaluation in Anlehnung an das kommerzielle ETL-Tool Survey. Diese Evaluation zeigte, dass die sehr bekannten Open Source Tools PDI und TOS ähnlich überzeugen konnten, wie der OWB. Es ist daher interessant zu verfolgen wie die dahinter stehenden Firmen diesen Erfolg ihrer Community Editions auch in Zukunft zu ihrem Vorteil werden nutzen können und ob die noch bestehende Lücke zum OWB ganz geschlossen werden kann.

Kontaktadresse:

Prof. Dr. Reinhold von Schwerin
Hochschule Ulm, Fakultät für Informatik
Prittwitzstr. 10
D-89075 Ulm

Telefon: +49(0)731-5028259
E-Mail: r.schwerin@hs-ulm.de
Internet: www.hs-ulm.de/r.schwerin

Quellen

[1] DELL: *DELL DVDStore*. http://linux.dell.com/dvdstore. Version: 13.12.2008

[2] div. Vorlesungsteilnehmer Sommersemester 2011, Wirtschaftsinformatik, Hochschule Ulm: *A comparison of an OS ETL-Tool with Oracle Warehouse Builder*. 2011. – Prüfungsvorleistung für das Fach Data Warehousing

[3] Kemper, Hans-Georg ; Mehanna., Walid ; Unger, Carsten: *Business Intelligence – Grundlagen und praktische Anwendungen*. 2. Aufl. ViewegTeubner, 2006

[4] Schwerin, Reinhold von: SQL/OLAP in der Lehre. In: *DOAG Konferenz 2008*, 2008