Evaluierung von Data Warehouse-Werkzeugen

Hong Hai Do[1], **Thomas Stöhr**[1], **Erhard Rahm**[1], **Robert Müller**[1], **Gernot Dern**[2]

[1]Institut für Informatik, Universität Leipzig
Augustusplatz 10/11, 04109 Leipzig

[2]R+V Allgemeine Versicherung AG
John-F.-Kennedy Str. 1, 65189 Wiesbaden

Tel.: 0341 / 97 32 306
Email: hong@informatik.uni-leipzig.de

Evaluierung von Data Warehouse-Werkzeugen

Zusammenfassung. *Die wachsende Bedeutung von Data Warehouse-Lösungen zur Entscheidungsunterstützung in großen Unternehmen hat zu einer unüberschaubaren Vielfalt von Software-Produkten geführt. Aktuelle Data Warehouse-Projekte zeigen, daß der Erfolg auch von der Wahl der passenden Werkzeuge für diese komplexe und kostenintensive Umgebung abhängt. Wir präsentieren eine Methode zur Evaluierung von Data Warehouse Tools, die eine Kombination aus Bewertung per Kriterienkatalog und detaillierten praktischen Tests umfaßt. Die Vorgehensweise ist im Rahmen von Projekten mit Industriepartnern erprobt und wird am Beispiel einer Evaluierung führender ETL-Werkzeuge demonstriert.*

1. Einleitung

Die wachsende Akzeptanz von Data Warehouses hat eine rasante Technologieentwicklung in diesem Bereich zur Folge. Dieser Markt verlangt nach unterstützenden Werkzeugen, die der Bereitstellung der notwendigen Informationen im Warehouse zum Entscheidungsprozeß dienen (Chaudhuri, Dayal 1997; Jarke et al. 2000; Kimball et al. 1998).

Abbildung 1 zeigt eine typische Data Warehouse-Umgebung. Daten aus heterogenen operativen Systemen (Dateien, DBMS[1], etc.) auf unterschiedlichen Plattformen (Mainframe, Client/Server, PC) müssen extrahiert, transformiert, aggregiert und schließlich in das zentrale Data Warehouse bzw. lokale Data Marts geladen werden (ETL[1] - Wieken 1999; Soeffky 1999). Typischerweise werden Data Warehouse bzw. Data Marts mittels CASE[1]-Produkten modelliert. Die Analyse der zusammengeführten Daten geschieht dann über BI[1]-Werkzeuge (z.B. OLAP[1], Reporting, Data Mining) der Datenzugriffskomponente (Wieken 1999). Zunehmende Bedeutung erfährt die Anbindung an das Internet und die Bereitstellung rollenspezifischer Informationsangebote über Informationsportale (White 1999). Metadaten, also beschreibende Informationen zu unterschiedlichsten Aspekten (Herkunft, Qualität, etc.) sollten in einer entsprechenden Management-Komponente (Repository) einheitlich und konsistent zur Nutzung und Administration des Data Warehouse verwaltet werden (Staudt 1999).

Zur Bewertung und Auswahl einer adäquaten Werkzeuglandschaft bedarf es daher einer strukturierten, praxisnahen und herstellerunabhängigen Vorgehensweise.

---

[1] BI = Business Intelligence, CASE = Computer Aided Software Engineering, DBMS = Datenbank-Managementsystem, ETL = Extraction Transformation Load, OLAP = Online Analytical Processing

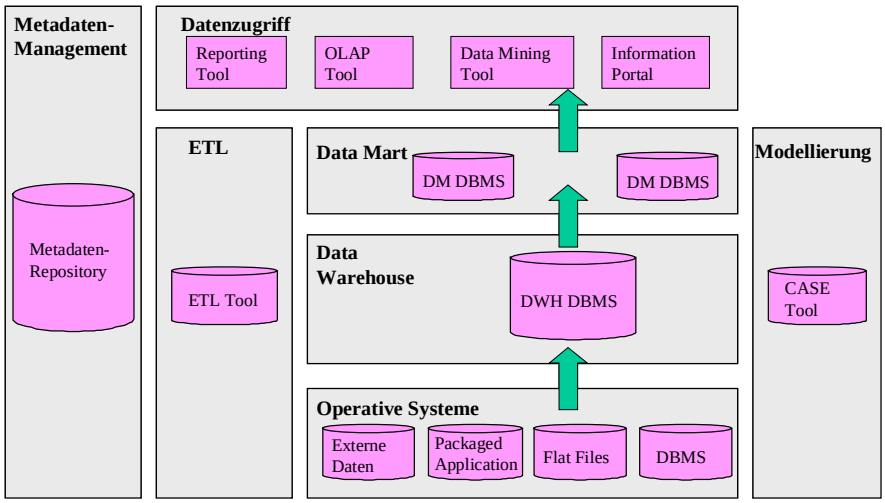

Abbildung 1. Data Warehouse-Architektur

Die Universität Leipzig (Abteilung Datenbanken) entwickelt u.a anbieterunabhängige Auswahlverfahren für verschiedene Klassen von Data Warehouse-Werkzeugen (ETL- und Portallösungen) und leitet daraus in Zusammenarbeit mit Unternehmen adäquate Produktvorschläge ab. Die erworbenen praktischen Erfahrungen auf diesem Gebiet resultieren aus bisher vier seit April 1998 durchgeführten Industrieprojekten mit der R+V Versicherung sowie dem Informatikzentrum der Sparkassenorganisation (SIZ). Zwei der Projekte konzentrierten sich auf die Evaluierung von ETL-Werkzeugen, um den Auswahlprozeß bei der Versicherung bzw. in den vom SIZ vertretenen Bankinstituten zu unterstützen (Müller et al. 1998; SIZ-Studie 1999).

Die praktische Evaluierung des State-of-the-Art von Data Warehouse-Werkzeugen ist für uns ein initialer Schritt zur Forschung in offenen Problemen dieses Bereiches. Neben dem Problem der Datenintegration (Härder et al. 1999; Jarke et al. 2000) betrifft dies vor allem den Bereich Metadaten-Management. Hier behandeln wir Ansätze zur Verbesserung der Interoperabilität sowie zur Bereitstellung von Business-Sichten auf Warehouse-Daten durch Integration von technischen und Business-Metadaten (Do, Rahm 2000; Müller et al. 1999).

Dieser Bericht zeigt eine in Industrieprojekten erprobte Vorgehensweise zur Auswahl von Data Warehouse-Produkten und präsentiert die daraus abgeleiteten Ergebnisse am Beispiel von ETL-Werkzeugen. Er ist wie folgt strukturiert: Kapitel 2 präsentiert unsere Vorgehensweise bei der Bewertung von Data Warehouse-Werkzeugen. Kapitel 3 vertieft dies am Beispiel der praktischen Evaluierung von ETL-Werkzeugen und präsentiert Ausschnitte unserer Produktergebnisse. In Kapitel 4 fassen wir zusammen und geben einen Ausblick auf geplante Tätigkeiten.

2. Vorgehensweise bei der Werkzeugevaluierung

Abbildung 2 verdeutlicht unsere im Evaluierungsprozeß verfolgte Vorgehensweise. Vier prinzipielle Schritte werden unterschieden:

- Projektspezifische Adaption des Evaluierungsverfahrens
- Anbietervorauswahl per Kriterienkatalog
- Praktische Evaluierung
- Bewertung der Testergebnisse und Empfehlung

Anschließend sollte eine produktionsnahe Erprobung unter Regie des jeweiligen Unternehmens erfolgen.

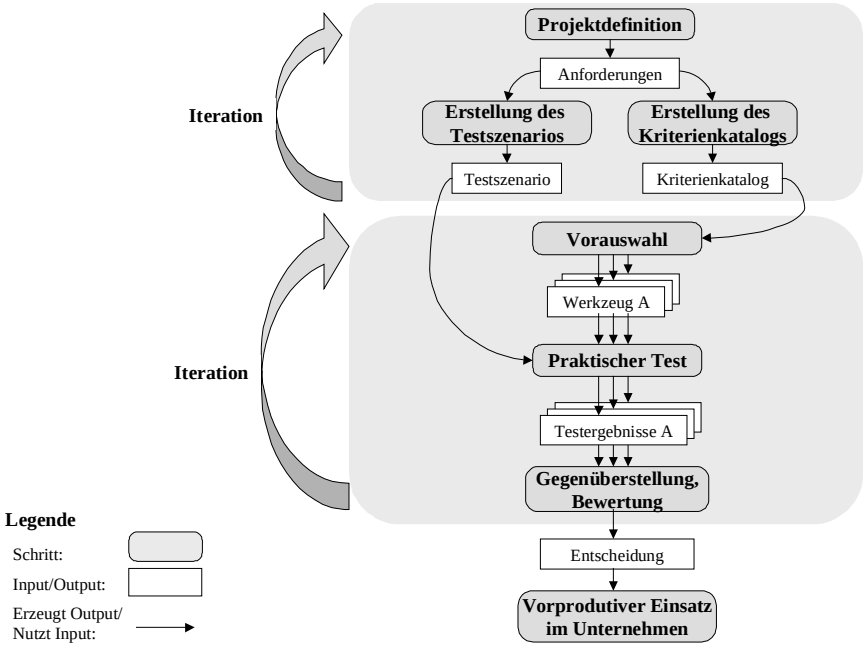

Abbildung 2. Vorgehensweise zur Werkzeugevaluierung

2.1. Projektspezifische Adaption des Evaluierungsverfahrens

Zunächst wird gemäß der unternehmensspezifischen Anforderungen die Klasse der benötigten Werkzeuge und die Anforderungen an diese im Rahmen des Data Warehouse-Projekts identifiziert. Insbesondere werden projektspezifische K.O.-Kriterien bestimmt (z.B. Minimalfunktionalitäten, technische Installationsvoraussetzungen, maximale Aufwendungen, etc.), die eine einfache Vorselektion von Produkten zu einem frühen Zeitpunkt ermöglichen.

Kriterienkatalog. Anschließend ist ein entsprechender Kriterienkatalog zur Vorauswahl der Werkzeuge zu entwickeln bzw. an die jeweiligen Projektanforderungen anzupassen. Diese Kriterien bilden die Basis für die vergleichende Bewertung. Um den Aufwand der Anbieter beim Ausfüllen zu minimieren, sollen die Kriterien möglichst eindeutig, gut kommentiert und vertretbar feingranular ausgewählt werden.

Testszenario. Die Papier-Evaluierung mittels Katalog (zusammen mit Informationen aus Data Sheets, White Papers, etc.) ergibt zwar bereits einen guten Überblick über die zu erwartende Mächtigkeit der Werkzeuge und ihre Unterschiede. Unsere Erfahrung hat allerdings gezeigt, daß erst der praktische Test vor Ort unter „Wettbewerbsbedingungen" die tatsächlichen Fähigkeiten der Werkzeuge aufzeigt. Daher wird ein einheitliches Testszenario benötigt, welches die zukünftige Einsatzumgebung der Werkzeuge im Unternehmen (Software, Hardware, Datenbasis) möglichst genau nachbildet und die Bewertung deren prinzipiellen Eigenschaften und Fähigkeiten in Form von praktisch zu lösenden Aufgaben adressiert.

Falls neue wesentliche Anforderungen in Betracht zu ziehen sind, ist ggf. eine Iteration dieser Phase erforderlich.

2.2. Vorauswahl per Kriterienkatalog

Per Recherche im Internet, in Fachzeitschriften und vergleichbaren Studien ist zunächst ein Überblick über den technologischen Stand des Marktes zu gewinnen und die Menge relevanter Anbieter und Produkte zu identifizieren.

Der Kriterienkatalog wird zunächst an in Frage kommende Anbieter versendet, welche den Kriterienkatalog innerhalb einer angemessenen Frist ausfüllen. Die anschließende Auswahl der Werkzeuge für den praktischen Test erfolgt dann über die Auswertung der ausgefüllten Kriterienkataloge. Mit einer Bewertungsmetrik durch Gewichtung von Kriterien gemäß Unternehmensanforderungen kann zu diesem Zeitpunkt aus den verbliebenen Werkzeugen bereits eine Vorauswahl getroffen werden, die den Kreis der ggf. aufwendig zu testenden Tools signifikant einschränkt. Insbesondere können Anbieter durch Erfüllung von K.O.-Kriterien frühzeitig ausgesondert werden.

2.3. Praktische Evaluierung

Die nach Durchführung der Katalog-Evaluierung gewonnene Einschätzung der Fähigkeiten der Werkzeuge durch Anbieter und Tester/Anwender können in einigen Punkten auseinandergehen bzw. Aspekte noch unklar bleiben. Außerdem sind Bewertungen über Ergonomie, Handling, Nutzerfreundlichkeit, etc. als wesentlich einzustufen.

Daher wurden praktische Tests gemeinsam mit den Anbietern in einem *Proof of Concept* (PoC) von zwei bis vier Tagen Dauer durchgeführt. Dadurch konnten die Installations- und Einarbeitungszeiten signifikant verkürzt werden. Gemäß eigenentwickelter Aufgabenstellungen wurden die Funktionalitäten mit dem Anbieter gemeinsam getestet und der Grad der Aufgabenbewältigung festgehalten.

2.4. Bewertung der Testergebnisse

Sämtliche Testergebnisse der Werkzeuge wurden in kompakten, gemäß der zentralen Bewertungskriterien strukturierten Testberichten zusammengefaßt. Die herausragenden Stärken und Schwächen der einzelnen Werkzeuge wurden mit einer +/-Bewertung verdeutlicht. Eine feinere Notengebung wurde nur in Ausnahmefällen, abhängig von den Projektzielen, vorgenommen, da die zugrundeliegende Gewichtung und Vergleichbarkeit generell als problematisch zu bewerten sind.

Sollte der Evaluierungs- und Entscheidungsprozeß länger andauern bzw. der Projektkontext wechseln, so ist ggf. eine Aktualisierung der gewonnenen Ergebnisse erforderlich. Bei verhältnismäßig kurzen Zeiträumen (z.B. ein Quartal) sind normalerweise keine wesentlichen Änderungen am Produkt zu erwarten.

Nach getroffener Auswahl sind die Kandidaten einer vorproduktiven Einsatzphase im Unternehmen zu unterziehen. In dieser Phase sollten, falls relevant, insbesondere Performance-, Verfügbarkeits- und Heterogenitätsaspekte der zu testenden Werkzeuge praktisch betrachtet werden.

3. Evaluierung von ETL-Werkzeugen

Der ETL-Vorgang wird in aktuellen Data Warehouse-Umgebungen sehr häufig durch eigenentwickelte Legacy-Programmierlösungen (C/C++, Cobol, DB-Skripte) durchgeführt. Insbesondere problematisch sind dabei

- der beträchtliche Aufwand bei der Wartung, Konsistenzerhaltung und Dokumentation,
- die schwierige Integration der Mainframe-Welt mit Client/Server- oder PC-Umgebungen und
- die fehlende Metadatenunterstützung.

In diesem Abschnitt wird die in Kapitel 2 beschriebene Vorgehensweise am praktischen Beispiel der Evaluierung von ETL-Werkzeugen diskutiert, welche im Kontext des Initialprojektes mit der R+V-Versicherung durchgeführt wurde. Ziel war die Identifikation eines adäquaten Metadaten-gestützten ETL-Werkzeuges.

3.1. Kriterienkatalog

Eine detaillierte Abhandlung des nachfolgend vorgestellten Kataloges findet sich in (Müller et al. 1998).

Anbieter-/Produktprofil. Anbieterinformationen wie z.B. Mitarbeiterzahl in Deutschland/weltweit, Marktpräsenz, Zeitpunkt Erst-Release, Vertriebsnetz, Bilanzsummen und nicht zuletzt der angebotene technische Support sind neben den Produktfähigkeiten wichtige Indikatoren für die Verläßlichkeit und Stabilität eines Anbieters (jüngst zahlreiche Übernahmen!).

Das Produkt wird grob gemäß der prinzipiellen Ausrichtung der Anbieter (z.B. DBMS-, BI-, CASE-, Repository-Bereich, etc.) klassifiziert. Außerdem werden die wesentlichen Komponenten und Funktionalitäten des Werkzeugs sowie die für die Installation und Nutzung notwendigen Hardware- und Software-Voraussetzungen registriert.

Quellextraktion/–integration. Ein zentraler Punkt ist die Fähigkeit zur Extraktion von Daten - bzw. bei knappem Zeitfenster für die Aktualisierung des Warehouse - auch von Änderungen auf Daten (*Changed Data Capture*) aus den relevanten Quellsystemen (diverse Dateiformate, relationale/hierarchische DBMS auf unterschiedlichsten Plattformen).

Insbesondere muß bei der Integrationsfähigkeit bewertet werden, inwiefern die Mapping-Definition sowie der Datentransfer von den Mainframe-Quellsystemen zu Client/Server- oder PC-Zielsystemen möglichst transparent und integriert, d.h. weitgehend ohne explizite Berücksichtigung der Systemgrenzen, durchgeführt werden kann.

Mapping-Definition, Datenbewegung. Zur Umsetzung der Abbildung von operativen in dispositive Systeme werden sog. *Mappings* definiert, welche die notwendigen Transformations- und Filtervorschriften enthalten. Deren Definition kann unterschiedlich erfolgen: graphisch oder skriptorientiert, per Drag&Drop oder tabellen-/formularorientiert. Es ist zu untersuchen, ob und bis zu welchem Komplexitätsgrad Filterbedingungen für das Lesen der Quellsysteme und Konvertierungsregeln bzgl. der Datenbewegungen definiert werden können (s. Abschnitt 3.2). Einen diesbezüglichen Ausschnitt unseres Kriterienkataloges zeigt Abbildung 3.

Bzgl. der Umsetzung der Datenbewegung finden sich prinzipiell zwei Ansätze: einige Produkte generieren Programm-Code (C/C++, Cobol, etc.), welcher dann lokal kompiliert wird und die Datenbewegungen durchführt. Bei der zweiten Variante, der sog. engine-basierten Lösung, wird der generierte Transformations-Code (im folgenden *Engine* genannt) zur Laufzeit interpretiert. Die Programme müssen in beiden Fällen in proprietäre oder externe Scheduling-Systeme eingebunden werden, welche wiederum verschiedene Grade an automatischer Verwaltung dieser Transformations-Jobs anbieten.

Abbildung graphisch definierbar				z.B. Verknüpfung von Quell- und Zielattributen durch Mausoperationen
Abbildung durch nicht-graphische (z.B. regelbasierte) „Skript"sprache definierbar				
Versionsmanagement der Abbildungsdefinitionen				MD-Produkt kann über frühere Versionen der Abbildungen Auskunft geben
n:m-Abbildungen zwischen Quell/Ziel-Objekten möglich				Objekt = **Tabelle** oder **sequentielle Datei**
1:n-Abbildungen zw. Quelle/Ziel-Attributen möglich				z.B. Quell-Attribut NAME (im Format „Müller, Stefan") kann in zwei Ziel-Attribute VORNAME („Stefan") und NACHNAME („Müller") zerlegt werden
n:1-Abbildungen zw. Quelle/Ziel-Attributen möglich				Analog zu 1:n-Abbildungen
(Teil-)Automatische Erkennung von Homonymen und Synonymen				z.B. aufgrund gleicher Datentypen und Wertebereiche sowie „ähnlicher" Namen
Kapazitäten/Möglichkeiten bei Data Clean-Up Rules				z.B. Quell- und Zielattribut (z.B. für ZAHLUNGS-MODUS) haben den gleichen Datentyp INTEGER, benutzen aber unterschiedliche Codes z.B. für die monatliche Zahlungsweise. Somit muß z.B. der Code-Wert 12 des Quellattributs auf den Code-Wert 0 im Zielsystem konvertiert werden.
Weitere Angaben zu deklarativen Abbildungsdefinitionen ...				

Abbildung 3. Ausschnitt des Kriterienkatalogs zum Aspekt Mapping-Komplexität

Metadaten-Management. Die Verwaltung von im ETL-Prozeß anfallenden Metadaten erfordert i.d.R. die Existenz eines Repositories. Hier ist zu unterscheiden, ob Produkte auf bewährte DBMS-Technik zurückgreifen oder eine einfache Dateiablage bevorzugen.

Für eine integriertes Warehouse-Konzept spielt die Interoperabilitäts– und Austauschfähigkeit eine tragende Rolle. Es werden Export/Import-Formate (MDIS[2], CDIF[3], XML[4] etc.) abgefragt, API-Fähigkeiten (C/C++ oder Java), toolspezifische Metadaten-Wrappers für BI-, Repository-Produkten, Portalen, etc. registriert. Weiterhin wird die Unterstützung adäquater Metamodellstandards (CWM[5], OIM[6]) vermerkt.

Neben den rein technischen Metadatenaspekten ist die Orientierung eines ETL-Produktes hinsichtlich Business Term-Modellierung, Grad der Komplexität des Metamodells (Entity-Relationship, UML[7]) bzw. die Art der Spezifikation des Business Term-Modells (graphisch, skriptorientiert) von Bedeutung.

Ergonomie, Handling. In diesem Abschnitt werden ergonomische Aspekte erfragt, z.B. die Existenz graphischer Benutzeroberflächen, kontextsenstiver Online-Hilfe, Unterstützung bei der Mapping-Definition, Komplexität der Administration, etc. Diese Kriterien können nur bedingt durch den Anbieter ausgefüllt werden;

[2] Metadata Interchange Specification – Metadata Coalition (MDC)
[3] CASE Data Interchange Format – Electronic Industries Alliance (EIA) und International Standards Organization (ISO)
[4] Extensible Markup Language – World Wide Web Consortium (W3C)
[5] Common Warehouse Metamodel – Object Management Group (OMG)
[6] Open Information Model – MDC
[7] Unified Modeling Language – OMG

i.d.R. ergeben sich diesbezügliche Ergebnisse erst im praktischen Umgang mit dem Werkzeug.

3.2. Praktisches Testszenario

Für die praktischen Tests steht eine Reihe verschiedenster Plattformen und kommerzieller DBMS (u.a. DB2, ORACLE, INFORMIX, SQLSERVER) zur Verfügung. Quell- und Zielsysteme für die zu simulierenden Datenbewegungen werden auf unterschiedlichen Plattformen/Betriebssystemen (Unix, NT) in unterschiedlichen DBMS eingerichtet, um den Data Warehouse-inhärenten Heterogenitätsanforderungen gerecht zu werden. Bei der Wahl der Plattformen/DBMS wird eine maximal mögliche Überlappung mit der Umgebung der jeweiligen Projektpartner angestrebt. Für die Datenbewegungstests wurden entsprechende Datenbasen und Testfälle generiert. Diese umfassen verschiedenste Aufgabenstellungen, auf deren Details wir aber aus Platzgründen nicht näher eingehen werden:

- Installation der Produktsuite (Bewertung der Komplexität und Fehleranfälligkeit dieses Vorganges)
- Einlesen von Schemainformationen aus Dateien und Datenbankkatalogen in das Repository
- Lösung komplexer Mapping- und Transformationsaufgaben. Es wurden die im realen ETL-Prozeß üblichen Abbildungstypen zwischen ein oder mehreren[8] Quell- bzw. Zielobjekten (Datenbanktabellen bzw. Flat Files) getestet
- Einbindung externer Funktionen in die Mapping-Definition
- Durchführen von Datenbewegungen
- Einfluß von Schemaänderungen usw.

3.3. Werkzeugtests und Ergebnisse

Bisher wurden die ETL-Werkzeuge DECISIONBASE (CA/PLATINUM), EXTRACT (ETI), POWERMART (INFORMATICA), COPYMANAGER (INFORMATION BUILDERS), DATASTAGE (INFORMIX/ARDENT), METASUITE (MINERVA SOFTCARE/CARLETON) und WAREHOUSEADMINISTRATOR (SAS) getestet und bewertet.

Im Kontext des Projektes mit der R+V-Versicherung wurden als finale Kandidaten DATASTAGE, POWERMART und COPYMANAGER identifiziert. Eine Hauptrolle spielte hierbei die bei einer Engine-Lösung trotzdem zu erwartende Mainframe-Anbindung. Als Ergebnis des SIZ-Projekts wurden die bis dato erstellten Testbe-

[8] je nach Kardinalität der Abbildung (Anzahl Quellen, Ziele) unterscheiden wir in der Notation entsprechend 1:1-, 1:n-, n:1- und n:m-Abbildungen

richte (alle außer desjenigen über IBI[9]) und für den Projektkontext adaptierte Bewertungskriterien in die entstandene Studie aufgenommen.

Tabelle 1 zeigt einen grobgranularen Überblick der Testergebnisse für die Werkzeuge COPYMANAGER, DATASTAGE und POWERMART mit Stand 10/99. Aus Platzgründen können wir hier nur einen kleinen Ausschnitt unserer umfangreichen Testergebnisse über einer Teilmenge der Bewertungskriterien sowie eine verkürzte Diskussion der Ergebnisse präsentieren (Müller et al. 1998; SIZ-Studie 1999).

Aspekt		**COPYMANAGER 4.2.1 (IBI)**	**DATASTAGE 3.5 (ARDENT)**	**POWERMART 4.5 (INFORMATICA)**
Anbieterprofil	#Mitarbeiter (weltweit/D)	1900/20	750/20	200/6
Produktprofil	Architektur	Client/Server		
	Installation (Server)	lauffähig auf NT/Unix/MVS	lauffähig auf NT/Unix	lauffähig auf NT/Unix
Mainframe-Integration	Kopplungsmechanismus	Installation der Treiber-Software auf Quellplattformen	Integration DATASTAGE /WAREHOUSEEXECUTIVE geplant	PowerConnect für DB2
Mapping-Funktionalität	Mapping-Design	tabellen-/formularorientiert	Drag&Drop	Drag&Drop
	Komplexität	Siehe Detailergebnisse in Tabelle 2		
	Datenbewegungsansatz	Engine	Engine (NT/Unix), Codegenerierung (MVS)	Engine
	Scheduler	ja		
	Ausführung von Job-Gruppen	hierarchische Abhängigkeiten zw. Jobs möglich	komplexe Abhängigkeiten zw. Jobs möglich	hierachische Abhängigkeiten zw. Jobs möglich
Metadaten-Management	Art der Verwaltung (Repository)	Dateien + RDBMS	RDBMS	RDBMS
	Versionierung	-	-	ja
	Impact Analysis	-	-	Anfragen auf Metadaten
	Interoperabilität	Siehe Detailergebnisse in Tabelle 3		
Nutzerfreundlichkeit	notwendige Programmierkenntnisse	SQL92 + proprietäres 4GL	SQL-Dialekte + proprietäres BASIC	SQL-Dialekte

Tabelle 1. Überblick Werkzeug-Gegenüberstellung (Ausschnitt)

Produktprofil. Alle getesteten Werkzeuge verfolgen einen Client/Server-Ansatz. Der Server übernimmt die Datenbewegungen und die Repository-Verwaltung, die Clients den Import von Schemainformationen, die Definition von Mappings und

[9] IBI war zum Zeitpunkt der Erstellung der SIZ-Studie noch nicht evaluiert

das Starten und Überwachen der assoziierten Datenbewegungen. Die überwiegende Anzahl der Werkzeuge verfolgt dabei den Engine-Ansatz (außer: ETI, METASUITE, Mainframe-seitig DATASTAGE).

Mainframe-Integration. Die Anbindung der mainframe-basierten Datenhaltungssysteme stellt immer noch eine Hürde speziell für engine-basierte Werkzeuge dar bzw. ist nicht vollzogen. Üblicherweise ist der Interpreter nur auf Windows NT- oder Unix-Plattformen lauffähig. COPYMANAGER ist derzeit der einzige Vertreter, welcher mit auf Quellplattformen verteilt installierter Treiber-Software eine transparente Integration der Mainframe-Seite ermöglichen kann. Durch die Kopplungsfähigkeit der DATASTAGE-Engines mit den WAREHOUSEEXECUTIVE-Programmen auf Mainframe-Seite (ehemals PRISM) besteht grundsätzlich die Möglichkeit zur Integration auch bei DATASTAGE, ist jedoch nach unserem Kenntnisstand derzeit nicht transparent umgesetzt. POWERMART bietet z.Z. nur geringe diesbezügliche Unterstützung (nur für DB2-Quellen) an.

Mapping-Komplexität[8]	**COPYMANAGER 4.2.1 (IBI)**	**DATASTAGE 3.5 (ARDENT)**	**POWERMART 4.5 (INFORMATICA)**
1 : 1	(ja), aber keine automatische Datentypkonvertierung DB2 *Date* → INFORMIX *Datetime*	ja	
1 : n	(ja) allerdings:		
n : 1	• Split in mehrere Sub-Mappings notwendig	(ja) allerdings:	(ja) allerdings:
n : m		• Kenntnisse über SQL-Syntax des Quellsystems nötig • Join[10] zwischen heterogenen Quellen ist nur zwischen Flat Files möglich! (Tabellen sind daher zunächst in Flat Files abzubilden)	• Kenntnisse über SQL-Syntax des Quellsystems nötig • explizite Umbenennung von Tabellen-Aliasnamen in SQL-Anfragen bei Recursive-Join[11]

Tabelle 2. Ausschnitt der Detailergebnisse zum Erfüllungsgrad von Mapping-Tests

Mapping-Definition/-Durchführung. Die Mapping-Definition erfolgt bei allen Anbietern graphisch. Eine nutzerfreundlichere Variante per Drag&Drop Mapping-Definition bieten fast alle Tools, lediglich EXTRACT, COPYMANAGER und METASUITE begnügen sich mit formularorientierter Darstellung. Tabelle 2 zeigt einen Ausschnitt unsere praktischen Ergebnisse der bewältigten Komplexität im Rahmen der Mapping-Tests. Die gestellten Mapping-Aufgaben konnten zwar prinzipiell bewältigt werden, jedoch mit äußerst unterschiedlichem Aufwand (mehrstufige Mapping-Definition möglich/nicht möglich, hoher/niedriger Pro-

[10] Hier ist die übliche Verknüpfungsoperation (=, >, <, ..) zwischen Attributen gemeint, wie sie z.B. in Sprachen wie SQL genutzt wird

[11] Join einer Tabelle mit sich selbst

grammieraufwand usw., Einschränkungen bei Verknüpfungen heterogener Quellen, etc.)

Alle Werkzeuge (außer ETI und METASUITE, da keine Engine) verfügen auch über einen eigenen Scheduler zum Anstoßen von Datenbewegungs-Jobs. Mit COPYMANAGER können zusammenhängende Job-Gruppen lediglich sequentiell abgearbeitet werden. Dagegen verfügt DATASTAGE über eine Job-Kontrollsprache, welche die Definition von komplexen Ausführungsabhängigkeiten der Mapping-Jobs erlaubt. POWERMART ist in der Lage, Mapping-Jobs in hierarchisch angeordneten Batches zu organisieren. Innerhalb eines Batches werden die Sub-Batches bzw. die Jobs selbst entweder sequentiell oder parallel verarbeitet. Die Einbindung der erzeugten Engines in externe Scheduling-Systeme wird prinzipiell durch die auf Betriebssystemebene aufrufbaren Engine-Interpreter ermöglicht.

Metadaten-Management/-Interoperabilität. Für das Metadaten-Management nutzen die Werkzeuge überwiegend ein RDBMS-basiertes, zentrales Metadaten-Repository. COPYMANAGER allerdings speichert Schemametadaten in Dateien, Mapping-Metadaten in einer relationalen Datenbank. Automatische Versionierung von Metadaten wird derzeit von POWERMART, EXTRACT und DECISIONBASE unterstützt. Rudimentäre Impact Analysis-Funktionalität (Anfragen auf Metadaten bzw. Browsing in Metadatenabhängigkeiten) wird, außer von DATASTAGE und COPYMANAGER, bereits von allen Werkzeugen angeboten.

Metadaten-Interoperabilität		**COPYMANAGER 4.2.1 (IBI)**	**DATASTAGE 3.5 (ARDENT)**	**POWERMART 4.5 (INFORMATICA)**
Metamodellstandards		-	-	OIM
Dateiformat	Import	-	-	-
	Export		MDIS (WAREHOUSE-CONTROLCENTER)	
API	Import	-	-	proprietäres MX2
	Export	proprietäres EDA (METADIRECTORY)		
Metadata-Wrappers	Import	-	MetaBroker (ERWIN, ERSTUDIO)	PowerPlug (ERWIN, DESIGNER2000, POWERDESIGNER)
	Export		MetaBroker (IMPROMPTU, BUSINESS OBJECTS, METASTAGE)	Bridge (IMPROMPTU, BUSINESS OBJECTS)

Tabelle 3. Ausschnitt der Detailergebnisse zur Metadaten-Interoperabilität

Die Metamodelle CWM und OIM erfahren praktisch keine Unterstützung. Unter den getesteten Werkzeugen bietet nur POWERMART Unterstützung für OIM an. Aktuell ist der Metadatenaustausch in COPYMANAGER nur auf IBI's proprietäres Repository-Produkt METADIRECTORY beschränkt. DATASTAGE und POWERMART verfügen über reichere Schnittstellen (Austauschformate, API bzw. toolspezifische Metadaten-Wrappers), um mit Modellierungs-, Repository- und Datenzugriffswerkzeugen Metadaten auszutauschen (vgl. Tabelle 3). Die zahlreichen

grauen Felder verdeutlichen die verbesserungswürdige Situation im Metadatensektor.

Ergonomie, Nutzerfreundlichkeit. Alle Werkzeuge setzen SQL-Kenntnisse des Nutzers bei der Vorfilterung der Datenquellen voraus. Für die Definition von wiederverwendbaren Transformationen muß sich der Nutzer zusätzlich Kenntnisse über die proprietären Sprachen 4GL von COPYMANAGER bzw. BASIC von DATASTAGE aneignen. Einige Besonderheiten der Produkte wie das Synonym-Konzept in COPYMANAGER für eindeutige Alias-Namen der Datenquellen oder das Port-Konzept in POWERMART für Ein- und Ausgabeattribute sind gewöhnungsbedürftig.

4. Zusammenfassung und Ausblick

Wir haben eine Vorgehensweise zur Werkzeugevaluierung vorgestellt, welche eine Kombination einer kriterienkatalog-basierten Vorauswahl mit detaillierten praktischen Tests umfaßt.

Die praktischen Tests unter „Wettbewerbsbedingungen“ zeigen, das oftmals Wunsch und Wirklichkeit der Werkzeugfähigkeiten auseinandergehen. Exemplarisch ist im ETL-Kontext die Integration der Mainframe- und Client/Server-Welt sowie das Metadaten-Management zu nennen. Insbesondere bei der OS/390-Integration sind produktionsnahe Tests unabdingbar. Integrierte Metadaten-Management-Lösungen werden versprochen, treffen aber selten die wirklichen Anforderungen.

Data Warehouse-Projekte zeigen auch, daß die Nutzbarkeit durch Integration unterschiedlicher interner und externer Informationsquellen (insbesondere aus dem Internet, wie z.B. Mitbewerberinformationen) erhöht werden kann. Diese Informationen sollten personalisiert und browser-basiert in Informationsportalen zur Verfügung gestellt werden. Zudem werden mächtige, intelligente Suchlösungen sowie inhaltsorientiertes Auffinden, Aufbereiten und Bereitstellen von Informationen zusammen mit den BI-Instrumenten benötigt.

An der Universität Leipzig wurde in Zusammenarbeit mit der R+V-Versicherung das beschriebene Evaluierungsverfahren auf Informationsportale adaptiert. Portalprodukte mehrerer Anbieter wurden bereits getestet (CA/STERLING EUREKA:PORTAL, HYPERWAVE INFORMATION PORTAL, VIADOR E-PORTAL-SUITE) bzw. befinden sich aktuell im Testverfahren (SYBASE ENTERPRISE PORTAL).

Wesentliche Anforderung an eine Data Warehouse-Umgebung ist eine adäquate Metadatenlösung, welche vor allem Business-Konzepte anbietet und die heterogene Welt von den operativen Quellen bis zum Portalanwender möglichst integriert. Die in Industrieprojekten gewonnenen praktischen Erfahrungen bestätigten das weitgehende Fehlen akzeptabler Metadatenlösungen, was insbesondere die Nut-

zung von Business-Metadaten und deren Integration mit technischen Metadaten anbelangt (Stöhr et al. 1999). Tests von BI-Werkzeugen zeigten z.B., daß diese heute noch fast reine Metadaten-Insellösungen darstellen. Hier sind noch zahlreiche Forschungsprobleme zu lösen. Aktuell untersuchen wir Aspekte der Metadaten-Interoperabilität, Repository-Technologie, Schemaintegration und –mapping (Do, Rahm 2000; Müller et al. 1999).

Literatur

Chaudhuri, S., Dayal, U.: An Overview of Data Warehousing and OLAP Technology; In ACM SIGMOD Record, 26(1), 1997.

Do, H.H., Rahm, E.: On Metadata Interoperability in Data Warehouses; Techn. Bericht, Inst. für Informatik, Univ. Leipzig, März 2000. http://dol.uni-leipzig.de/pub/2000-13/

Härder, T., Sauter, G., Thomas, J.: The Intrinsic Problems of Structural Heterogeneity and an Appoach to Their Solution; In VLDB Journal (8), 1999.

Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P.: Fundamentals of Data Warehouses; Springer, 2000.

Kimball, R., Reeves, L., Ross, M., Thornthwaite, W.: The Data Warehouse Lifecycle Toolkit; Wiley Computer Publishing, 1998.

Müller, R., Stöhr, T., Rahm, E.: An Integrative and Uniform Model for Metadata Management in Data Warehousing Environments; Proc. Intl. Workshop on Design and Management of Data Warehouses (DMDW), 1999.

Müller, R., Stöhr, T., Rahm, E.; Quitzsch, S.: Evaluierung und Produktvorschlag für ein technisches Metadaten-Management im R+V-Data Warehouse; Projektbericht, Inst. für Informatik, Univ. Leipzig, November 1998.

SIZ-Studie: Data Warehouse Tools – DWT; Band I und II. Informatikzentrum der Sparkassenorganisation GmbH (SIZ) und Univ. Leipzig, Oktober 1999.

Soeffky, M.: Datenaufbereitung für das Data Warehouse; In it FOKUS (3), 1999.

Staudt, M., Vaduva, A., Vetterli, T.: The Role of Metadata for Data Warehousing; Techn. Report 99.06., Inst. für Informatik, Univ. Zürich, September 1999.

Stöhr, T., Do, H.H., Rahm, E.: Vergleich von Metadatenansätzen kommerzieller Anbieter für das Data Warehouse der R+V-Versicherung; Projektbericht, Inst. für Informatik, Univ. Leipzig, November 1999.

White, C.: Using Information Portals in the Enterprise; In Data Management Review, April 1999. http://www.dmreview.com/master.cfm?NavID=55&EdID=61.

Wieken, J.H.: Der Weg zum Data Warehouse; Addison-Wesley, 1999.