*i*ETL: Flexibilisierung der Datenintegration in Data Warehouses

Sebastian Schick[1], Gregor Buchholz[2], Meike Klettke[3], Andreas Heuer[1], Peter Forbrig[2]
[1]Lehrstuhl für Datenbank- und Informationssysteme; [2]Lehrstuhl für Softwaretechnik
[3]Institut für Informatik
Universität Rostock, 18051 Rostock
vorname.nachname@uni-rostock.de

ABSTRACT

Data Warehouses bestehen aus zwei Hauptkomponenten: einer flexiblen Anfrageschnittstelle zur Datenanalyse (OLAP) und einer relativ starren ETL-Komponente zum Laden der Daten ins Data Warehouse. In diesem Artikel soll vorgestellt werden, wie die Datenintegration bedarfsabhängig zu flexibilisieren ist, welche Vorteile sich daraus ergeben und welche Herausforderungen bei der Entwicklung eines solchen interaktiven ETL (*i*ETL)-Prozesses bestehen.

Categories and Subject Descriptors

D.2.2 [**Software Engineering**]: Design Tools and Techniques—*User interfaces*; H.2.7 [**Database Management**]: Database Administration—*data warehouse and repository*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*query formulation, search process*

General Terms

Data Warehouse

Keywords

Data Warehouse, ETL-Prozess, Szenario, Datenintegration

1. MOTIVATION

Institutionen des öffentlichen Sektors sehen sich mehr noch als industrielle Verwaltungen einer Vielzahl von Softwarelösungen zur Unterstützung ihrer Prozesse gegenüber. Während in Industrien wie dem Automobilbau oder dem Bankenbereich fünf bis zehn Kernkompetenzen in der IT dargestellt werden, finden sich im Kerngeschäft öffentlicher Verwaltungen leicht zwischen 100 und 150 Prozesse verschiedener Dienstleistungskomplexe [11]. Diese Aufgaben- und Informationsfülle spiegelt sich in der Heterogenität der vorzuhaltenden Lösungen, der Vielfalt der Schnittstellen zum Informationsaustausch sowie der Menge vorzufindender Datenbanklösungen und Datenablagen in Excel und ähnlichen Formaten wider. Dem gegenüber steht der zunehmende Bedarf an zentralen Beobachtungs- und Steuerungsinstrumenten der Bereiche Business- und Geo-Intelligence. Der Verbreitung fachübergreifender Systeme steht oft der sehr hohe Datenbeschaffungsaufwand (sowohl initial als auch prozessbegleitend) im Weg. Insbesondere die Erschließung neuer Datenquellen bei der Ausweitung von Kennzahlensystemen auf neue Fachgebiete oder bei Auftreten tagesaktueller *ad-hoc*-Abfragen mit teils „exotischen“ Fragestellungen geht stets mit großem manuellen Aufwand bei der Informationssuche und -transformation einher. Es geht also um eine Lösung zur Identifizierung und Integration heterogener Datenquellen im Data Warehouse (DW)-Umfeld, die den Anwender in verschiedensten Datenquellen enthaltene Informationen finden und in seinen Bestand übernehmen lässt. Nicht im Fokus stehen von technischem Personal eingerichtete turnusmäßige Beladungen oder Real-Time-Data-Warehousing sondern die zunächst einmalige Integration aus Quellen mit überwiegend statischem Inhalt und geringer Komplexität durch den Anwender. Kapitel 2 illustriert dies anhand zweier Szenarien aus der Anforderungsanalyse. Abb. 1 zeigt in durchgezogenen Pfeilen bestehende Daten- und Kontrollflüsse und in unterbrochenen Linien die zu entwickelnden Verbindungen. Interessant dabei ist, wie der Prozess aus Nutzersicht verlaufen kann und mehr noch, mittels welcher Architekturen und Datenstrukturen die daraus ermittelten Anforderungen am besten umzusetzen sind.

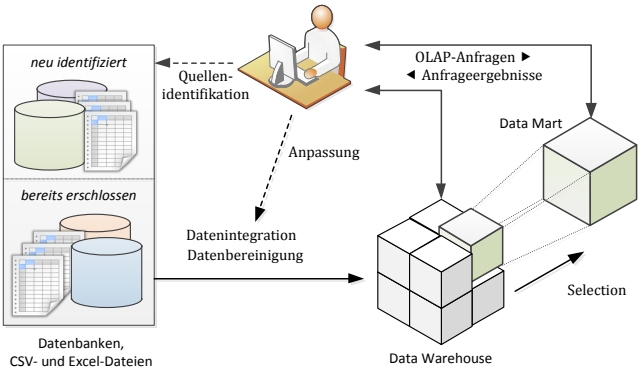

Abbildung 1: Interaktiver ETL-Prozess eines DW

*Die Arbeit dieser Autoren wird durch das BMWi im ZIM-Projekt KF2604606LF1 der GeoWare GmbH und der Universität Rostock gefördert.

24^{th} GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 29.05.2012 - 01.06.2012, Lübbenau, Germany.

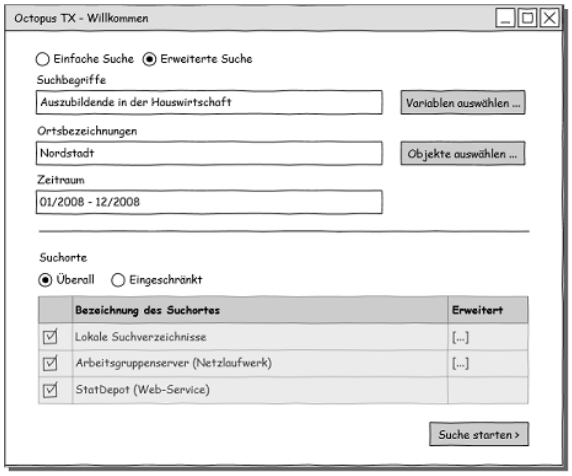

Abbildung 2: Eingabemaske Szenario 1

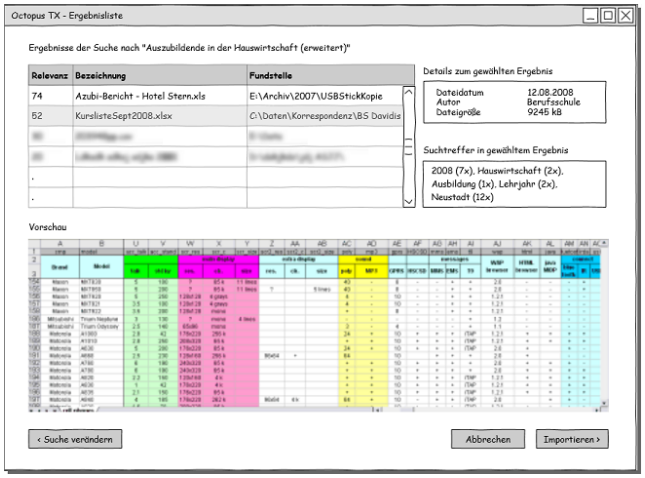

Abbildung 3: Ergebnisliste Szenario 1

2. ANWENDUNGSSZENARIEN

In der Anforderungsanalyse dieses Projektes sind Szenarien ([3], S. 52) zur Veranschaulichung der gewünschten Funktionalität entstanden, die beim Entwickeln einer Lösung helfen sollen. Die folgende Wiedergabe dieser Beispielanwendungen beginnt nach dem Skizzieren des technischen Kontextes jeweils mit der Beschreibung einer Bedarfssituation, an die sich eine mögliche Lösung aus Anwendersicht anschließt. Das folgende Kapitel 3 schlägt dann ein Konzept zur Umsetzung dieser Anforderung vor.

Szenario 1 – Kontext: Die Klassifikationshierarchie des DW mit ihren Attributen ist dem System bekannt. Ebenso wurden die möglichen Datenquellen (Web-Services, Suchpfade im Dateisystem) bereits konfiguriert. Die drei Dimensionen der Daten im DW sind: *Zeitraum*, *Geo-Objekte* (Hierarchie geographischer Bezugselemente) und *Kenngrößen*.

Situation: Wenige Tage vor der Jahresversammlung des Gaststätten- und Hotelverbandes wird Herr B. im Amt für Ausbildungsförderung mit dem Zusammenstellen einer Statistik beauftragt. Sie soll die Entwicklung der Auszubildendenzahlen der vergangenen Jahre in diesem Bereich aufzeigen. Dazu fragt er die dafür relevanten Informationen in einem DW-System an und erkennt an der grauen Einfärbung des entsprechenden Knotens in seiner DW-Anwendung, dass die Daten zur Kenngröße „Auszubildende in der Hauswirtschaft“ für den Stadtteil „Nordstadt“ im Jahr 2008 fehlen. Dass sie nicht wie sonst automatisch übermittelt wurden, liegt seiner Meinung nach an der Umbenennung des Stadtteils (früher: „Neustadt“) im Vorjahr.

Lösungsausblick: Über einen Rechtsklick auf den ausgegrauten Knoten lässt Herr B. eine Anfrage an das Suchsystem generieren, was ihn zur Eingabemaske (Abb. 2) führt. Die Suchfelder für die drei Dimensionen sind schon vorbelegt; per direkter Texteingabe oder über die Schaltflächen „Variablen auswählen“ und „Objekte auswählen“ könnte Herr B. die Suchkriterien verändern; die jeweiligen Dialoge stellen die möglichen Werte der jeweiligen Dimension zur Auswahl. Er tippt in das Suchfeld der Ortsbezeichnungen „Neustadt“ ein und bekommt nach Abschluss der Suche als Ergebnis eine Liste von Datenquellen zu sehen (Abb. 3). Das erste Ergebnis ist offenbar ein Bericht eines konkreten Hotels und damit nicht relevant. Herr B. wählt also den zweiten Treffer und betrachtet ihn in der Voransicht. Er erkennt, dass die gesuchten Daten (eine Auflistung der Auszubildenen mit Wohn- und Ausbildungsadresse) in dem Suchergebnis enthalten sind und wechselt via „Importieren“ zum Import-Modul für Excel-Dateien. Dort kann er einzelne Spalten und Bereiche der Excel-Tabelle als Werte der drei Dimensionen markieren und den Import in sein DW-Systems anstoßen. Anschließend widmet er sich der Aufbereitung der Statistiken.

Szenario 2 – Kontext: Die Konfiguration entspricht der von Szenario 1. Hier wird jedoch die Anfrage nicht vom DW-System vorbelegt.

Situation: Vor dem Beitritt zu einem Verband zur Förderung von Solarenergieanlagen an privaten Immobilien sollen für 2010 Anzahl, Verteilung und Höhe der Landesförderung von Anlagen ermittelt werden. Frau B. ist damit beauftragt und stellt fest, dass zu den Förderungen bislang keine Daten im DW existieren, jedoch hat sie kürzlich davon gehört, dass ein Mitarbeiter einem anderen per Mail eine solche Übersicht schicken wollte.

Lösungsausblick: Sie startet das Suchsystem und spezifiziert ihre Anfrage: „solaranlagen 2010 +förderung +privat -gewerblich“ (siehe Abb. 4). Der Begriff „solaranlagen“ und das Jahr „2010“ sollen im Ergebnis enthalten sein. Ebenso „förderung“ und „privat“, die ihr besonders wichtig sind und für die das „+“ eine erhöhte Priorität der Fundstellen bewirken soll. Kommt hingegen „gewerblich“ in der Fundstelle vor, soll die Priorität des Ergebnisses sinken. Als Suchort schließt Frau B. die Datenquelle „StarDepot“ aus, dann startet sie die Suche. In einer Ansicht ähnlich zu Abb. 3 nutzt sie die Vorschau, um die Datei mit den gesuchten Informationen zu identifizieren. Anschließend bereitet sie die Daten für den Import vor und übernimmt sie in den eigenen Datenbestand.

3. LÖSUNGSKONZEPT

In Abschnitt 2 wurden Anwendungsszenarien und mögliche Lösungsausblicke vorgestellt, die eine Anpassung des ETL-Prozesses notwendig machen. In diesem Abschnitt stellen wir dafür eine erweiterte DW-Architektur vor.

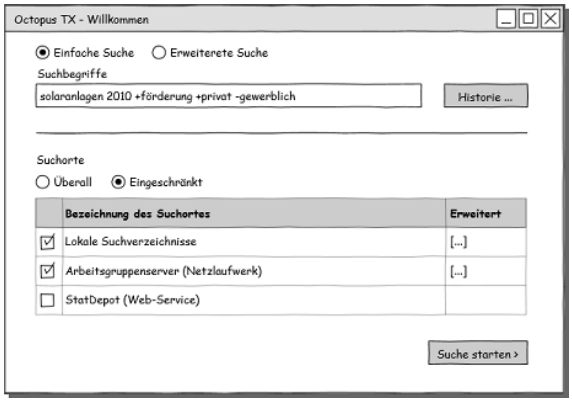

Abbildung 4: Eingabemaske Szenario 2

3.1 Ausgangspunkt

Der Anwender soll bei der Quellenidentifikation und Datenintegration im ETL-Prozess in einem DW unterstützt werden. Dafür müssen die verfügbaren Datenquellen so aufbereitet werden, dass eine Recherche und eine anschließende Auswahl von geeigneten Datenquellen möglich ist. Die Heterogenität der Datenquellen erschwert die automatische Integration in das DW. In föderierten Datenbanken gab es umfangreiche Untersuchungen zu Heterogenitäten der einzelnen Datenquellen bzgl. Syntax und Semantik von Werten, Attributen, Relationen und Modellen [6]. Die Transformation von Daten aus heterogenen Formaten in eine einheitliche Repräsentationsform stellt das Hauptproblem bei der Integration dar. Der Anwender muss deshalb bei der Datenintegration und insbesondere bei der Datentransformation unterstützt werden. Angepasste Nutzerinterfaces sollen den technikunerfahrenen Anwender unterstützen.

Der Prozess des Füllens eines DW mit Daten wird als ETL-Prozess bezeichnet, ETL steht hierbei für Extract, Transform und Load. Die Basisdaten in den meisten Anwendungen sind heterogene Daten, die in ein einheitliches Format, das multidimensionale Modell des DW integriert werden sollen. Bei diesem Prozess werden die neuen oder veränderten Daten aus den Basisdatenquellen ausgewählt (**E**xtraction), in eine einheitliche Darstellung umgewandelt (**T**ransformation), dabei vervollständigt, Duplikate bereinigt und eventuell voraggregiert. Anschließend erfolgt das Laden in die Datenbank (**L**oad) [5]. Im vorgestellten Ansatz soll eine Wissenskomponente den ETL-Prozess unterstützen, indem einzelne Komponenten um semantische und ontologische Konzepte erweitert werden. Wir schlagen deshalb eine Erweiterung des klassischen ETL-Prozesses in folgenden Bereichen vor:

- **Quellenidentifikation**: Methoden des Information-Retrieval sollen den Anwender bei der Identifikation und Vorauswahl von Datenquellen unterstützen.
- **Datenintegration**: Die flexible Integration heterogener Datenquellen soll durch semiautomatische Techniken gefördert werden.
- **Datenextraktion** (als Teil der Datenintegration): Der Anwender soll durch geeignete Nutzerinterfaces die Abbildungs- und Transformationsvorschriften effizient bestimmen können.
- **Wissensbasis**: Sämtliche Komponenten sollen zur Flexibilisierung um semantische und ontologische Konzepte erweitert werden.

Wir schlagen deshalb vor, das Referenzmodell für die Architektur von DW aus [2] derart zu erweitern, dass der Anwender bei der Identifikation passender Datenquellen unterstützt, der Integrationsprozess heterogener Datenquellen erleichtert und die Flexibilisierung der Datenextraktion mit geeigneten Konzepten ermöglicht wird. Die Architektur ist in Abbildung 5 dargestellt. Datenflüsse zwischen den Komponenten sind als durchgezogene Pfeile umgesetzt, der Kontrollfluss wird mit unterbrochenen Linien markiert.

3.2 Die Wissenskomponente

Die zentrale Komponente in der vorgestellten Architektur bildet der Data Warehouse Manager (DWM) (siehe Abb. 5). Der DWM steuert in einem klassischen DW nach [2] alle Komponenten, die zur Anfrage und Darstellung der Daten notwendig sind: Monitore, Extraktoren, Ladekomponenten und Analysekomponenten. Zusätzlich erhält der DWM in der hier vorgestellten Architektur Schnittstellen

- zur zentralen Wissenskomponente, die für die Planung und Ausführung der Quellenidentifikation und Datentransformation im ETL-Prozess benötigt wird und
- zur Search Engine zwecks Quellenidentifikation.

Konzept.

Die Wissenskomponente (**knowledge component**) stellt Informationen über Klassifikations- und Dimensionshierachien, semantische Verknüpfungen und die Typisierung sowie Metadaten einzelner Attribute bereit (siehe Abb. 5). Das domänenspezifische Wissen wird durch Quellenangaben, Synonyme und Muster (Format- bzw. Modell-Pattern) ergänzt. Die Wissenskomponente ist für die Prozesse der Quellenidentifikation und Datenintegration neuer Datenquellen unabdingbar.

- Der **Metadata Manager** stellt eine Schnittstelle bereit, über die andere Komponenten Anfragen an die Wissensbasis und das Metadaten-Repository stellen und Antworten anfordern können.
- Die **Knowledge Base** beinhaltet ein Wissensmodell für die Speicherung und Verwaltung der semantischen und ontologischen Informationen.
- Das **Metadata Repository** beinhaltet alle weiteren Metadaten, die vom DWM benötigt werden.

Herausforderungen.

Die Umsetzung einer Wissenskomponente erfordert den Aufbau einer Wissensbasis zum Anwendungsgebiet des DW, womit je nach Anwendungsszenario ein hoher manueller Aufwand verbunden ist. Ein Teil der Wissensbasis kann aus den hierarchischen Klassifikationsattributen der Dimensionen des DW übernommen werden.

Zusätzlich zu diesen hierarchischen Informationen werden Wörterbücher benötigt, die die Verbindung zwischen Konzepten der Wissensbasis und Suchbegriffen herstellen. Diese Wörterbücher sind initial zur erstellen und sollen beim Einsatz des *i*ETL-Tools von einer lernenden Komponente erweitert und anpasst werden.

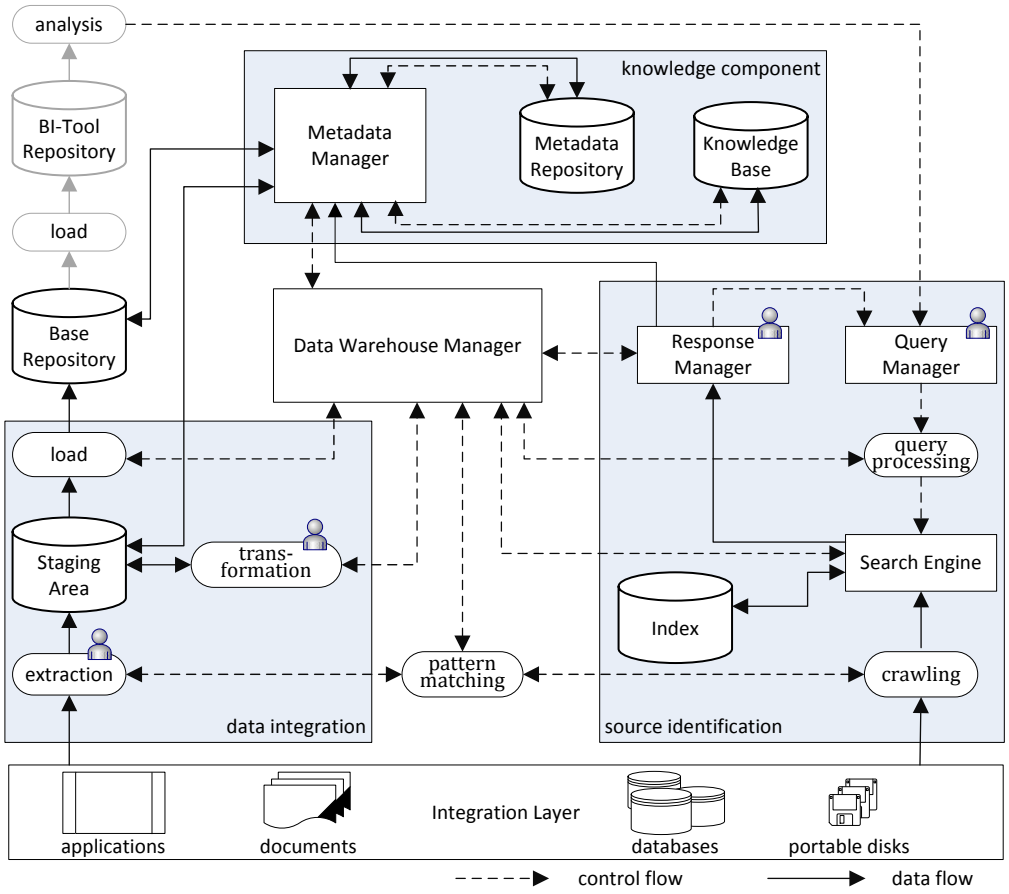

Abbildung 5: Architektur für den flexiblen ETL-Prozess

3.3 Quellenidentifikation

Liegt in einem klassischen DW die Auswahl der Datenquellen bzw. der Quelldaten vorrangig bei den zuständigen Administratoren, so soll in diesem Ansatz der Anwender des DW in die Quellenidentifikation einbezogen und dabei unterstützt werden.

Konzept.

Eine Suchkomponente soll bei der Quellenidentifikation im ETL-Prozess dem Anwender die Auswahl weiterer Datenquellen erleichtern. Dafür soll ein Index über alle verfügbaren Datenquelle aufgebaut werden, die über den **Integration Layer** verfügbar sind. Der Prozess der Quellenidentifikation (*source identification*) ist in Abbildung 5 rechts dargestellt. Die Komponenten sind an die Architektur einer Web-Suchmaschine angelehnt, wie sie beispielsweise in [1] (Seite 460) vorgeschlagen wird. Sie werden im Folgenden vorgestellt.

- Der **Query Manager** stellt über eine Schnittstelle einfache oder erweiterte Suchmasken bereit. Für die Suche werden vorerst die existierenden Dimensionen aus dem DW als Suchparameter verwendet:
  - *Was*: Kenngrößen, die in einer Datenquelle enthalten sein müssen.
  - *Wo*: Geo-Objekte, die durch die Werte beschrieben werden.
  - *Wann*: Datumsangaben und Zeitbereiche.
- Das **Query Processing** beschreibt eine Vorverarbeitung der Anfrage unter Verwendung der Wissensbasis. Dabei soll eine Anpassung der Anfrage hinsichtlich der Struktur (Ort, Zeit, Schlüsselwörter), die Expansion der Anfrage mit Hilfe der Wissensbasis sowie ein Vokabularmapping und weitere Vorverarbeitungsschritte wie eine lexikografische Analyse oder Stoppworteleminierung stattfinden.
- Die **Search Engine** soll die Indizierung von Datenquellen, eine Anfrageverarbeitung und die Bereitstellung von Ergebnislisten übernehmen. Die Indizierung (mit gleichen Methoden wie bei der Anfrageverarbeitung) soll durch eine Strukturanalyse auf Basis von Mapping-Mustern (die während der Datenintegration erzeugt werden) umgesetzt werden. Neben dem Bereitstellen von Anfrageergebnissen aus heterogenen Datenquellen sollen Suchergebnisse mit Informationen über den Fundort innerhalb der Datenquellen angereichert werden, wofür domänenspezifische Informationen und Musteranalysen aus der Wissensbasis zum Einsatz kommen sollen.
- Das **Crawling** beschreibt den Schritt, in dem verfügbare Datenquellen durchsucht und für die Indizierung bereitgestellt werden. In diesem Schritt ist die Nut-

zung vorhandener Mapping-Muster zu Strukturanalysen und semantischen Auswertungen geplant.

- Der **Response Manager** wird für die Präsentation der Anfrageergebnisse genutzt. Dem Nutzer soll eine Vorauswahl von Datenquellen durch eine vereinfachte Datenvorschau ermöglicht werden, eine semiautomatische Identifizierung möglicher Indikatoren, Variablen und Zeiträume in den Anfrageergebnissen soll dabei erfolgen (siehe Abb. 3). Die ausgewählten Quellen werden hier an den DWM übergeben, der in einem nächsten Schritt die Datenintegration anstoßen kann.

Herausforderung.

Eine Herausforderung ist das Design des Anfrageinterfaces und der Ergebnisdarstellung. Hierfür müssen die Anforderungen der Anwender bestimmt werden.

- **Anfrageinterfaces**: Wie können Suchanfragen auf Ordner, Datenquellen oder Dateitypen eingegrenzt werden und welche der Anfragetypen Context (Phrase, Boolean, etc.), Pattern Matching oder strukturierte Anfragen (formularbasiert) können genutzt werden. Außerdem ist zu klären, ob die Klassifizierung der Anfrage durch vorhandene Kategorien der Wissensbasis möglich und sinnvoll ist.

- **Ergebnisdarstellung**: Wie muss eine zielgerichtete Präsentation der Anfrageergebnisse unter Verwendung der Wissensbasis umgesetzt werden und wie ist dabei eine semiautomatische Identifizierung potentieller Indikatoren, Variablen und Zeiträume möglich. Daneben sollen relevante Elemente der Ergebnismenge hervorgehoben und die Identifizierung von Strukturen innerhalb eines Treffers durch Anwendung von Mapping-Mustern möglich sein.

- **Query Processing** (basierend auf der Wissensbasis): Wie kann die Wissensbasis für die Anfrageerweiterung in Form einer facettierten Suche genutzt und wie können Methoden des Relevance Feedback zur Verbesserung der Qualität der Ergebnisse genutzt werden.

- **Search Engine**: Wie kann die Integration externer Anwendungen (Enterprise Search, ERP, CRM, etc.) umgesetzt werden, wenn die bereitgestellte Anfrageschnittstellen nur Teilergebnisse liefern oder der Umfang der Datenbasen zu groß ist. Die Integration unterschiedlicher Datenformate soll ebenso unterstützt werden, wie die Duplikaterkennung, wenn Inhalte und Daten aus unterschiedlichen Quellen genutzt werden. Weiterhin sollen Mapping-Muster für den Prozess der Indizierung und Extraktion genutzt werden und Klassifikation durch die Mapping-Muster unterstützt werden.

3.4 Datenintegration

Die Datenintegration muss bedarfsabhängig und flexibel angepasst werden, wenn durch die Quellenidentifikation neue Datenquellen zu integrieren sind. Die Flexibilisierung soll durch Anwendung von semantischen und ontologischen Konzepten erreicht werden, wodurch domänenspezifisches Wissen ausgenutzt wird. Die Architektur in Abbildung 5 ist dabei an die Referenzarchitektur angelehnt.

Konzept.

- Mit **extraction** wird die Übertragung von Daten aus externen Quellen in den Arbeitsbereich (*staging area*) beschrieben. Die Auswahl der Datenquellen wurde im Vorfeld durch den Anwender (Quellenidentifikation) durchgeführt. Der Prozess muss um semiautomatische Methoden des Schema Matchings und Mappings erweitert werden (**pattern matching**).

  Die bei der Datenextraktion und -transformation erzeugten Mapping-Mustern sollen für eine spätere Wiederverwendung in der Wissensbasis vorgehalten werden und bei jeder Datenintegration auf ihre Anwendbarkeit hin überprüft werden. Passende Muster für die Extraktion einer Datenquellen werden dem Anwender angeboten. Die Schritte der Extraktion und Transformation müssen durch angepasste, graphische Tools unterstützt werden.

- Mit **transformation** wird die Abbildung der Daten hinsichtlich struktureller und inhaltlicher Aspekte beschrieben. Neben der Datentransformation (z. B. Konvertierung von Kodierungen, Vereinheitlichung von Datumsangaben, etc.) sollen hier auch eine Datenbereinigung, Duplikaterkennung und eine Datenfusion stattfinden (siehe auch [2]). Für diesen Schritt sind ebenfalls Informationen aus der Wissensbasis notwendig. Vorschläge für Transformationsvorschriften können aus der Wissensbasis abgeleitet werden.

- Mit **load** wird die Übertragung der Daten aus dem Arbeitsbereich in das **Base Repository** beschrieben. Die Daten stehen dann für die weitere Verarbeitung durch unterschiedliche BI-Tools (Business Intelligence Tools) zur Verfügung. Durch ein erneutes Laden werden die Daten in ein externes **BI-Tool Repository** geladen (grau hinterlegt) und stehen so in DW-Anwendungen für weitere OLAP-Analysen zur Verfügung.

Herausforderungen.

Die Herausforderungen bei der Datenintegration liegen bei der Tool-Unterstützung des Anwenders, sowie bei der semantischen Unterstützung des Transformationsprozesses. Das Schema einer Datenquelle ist in der Regel unbekannt, weshalb es mit Hilfe geeigneter Werkzeuge extrahiert werden muss.

- **Transformation**: Die Datentransformation aus dem Format der Basisdatenquelle ins Zielformat kann nicht vollständig automatisiert werden. Herausforderung ist hier die Entwicklung von Nutzerinterfaces zur Eingabe der benötigten Informationen durch den Fachanwender. Die dabei entstehenden Transformationsmuster sollen gespeichert werden, damit sie für andere Datenquellen verwendet werden können.

  Welche vorhandenen Ansätze der Datenintegration können für die Datenbereinigung, Duplikaterkennung und Datenfusion angewendet werden und wie kann eine Plausibilitätsprüfung der Daten unterstützt werden. Für eine Plausibilitätsprüfung können z. B. Regeln definiert werden, die die Wissensbasis einbeziehen. Ein möglicher Ansatzpunkt ist hier die Angabe von check-constraints.

3.5 Einsatz des Verfahrens

Das im Projekt zu entwickelnde Verfahren wird sich nicht auf alle DW anwenden lassen. Voraussetzung ist, dass es eine Wissensbasis zu dem Anwendungsgebiet des DW gibt. Da diese Wissensbasis eine zentrale Rolle beim Finden der relevanten Datenquellen und bei der Transformation der Daten ins DW spielt, muss eine solche Wissensbasis für einen flexiblen ETL-Prozess vorhanden sein. Teile der Wissensbasis lassen sich aus den Klassifikationsattributen der Dimensionen des DW generieren; die Zuordnung dieser Klassifikationshierarchie zu den korrespondierenden Suchbegriffen für die Datenquellen muss für das jeweilige Anwendungsgebiet ergänzt werden.

4. RELATED WORK / STAND DER TECHNIK

4.1 Datenintegration

Jede Datenintegration bewirkt das Zusammenführen von Daten aus heterogenen Datenbanken und Informationssystemen. Es gibt Klassifikationen, die die Heterogenitäten der einzelnen Datenquellen systematisieren. Heterogenitäten können bzgl. Syntax und Semantik von Werten, Attributen, Relationen, Modellen existieren ([6]). Eine Standardarchitektur, die das Zusammenführen von heterogenen Formaten in heterogenen Datenbanken vornimmt, wurde bereits im Jahr 1990 in [10] vorgeschlagen.

Eine dabei bestehende Aufgabe ist Matching und Mapping heterogener Datenbanken. Es gibt mehrere Mapping-Tools, die eine intuitiv bedienbare Oberfläche anbieten, um dem Benutzer das Entwickeln von Datentransformationskomponenten zu erleichtern (wie Altova MapForce[1], oder IBM Data Integrator), dieser Prozess ist jedoch nicht automatisierbar. Einen Überblick über Forschungsansätze in dieser Richtung findet man in [9]. Dabei spielen vor allem Ontologie-basierte Ansätze eine große Rolle (vgl. [7] und [4]).

4.2 ETL

Beim ETL-Prozess in einem DW werden die Basisdaten (meist heterogene Daten) in ein einheitliches Format, das multidimensionale Modell des DW integriert [5]. Man kann den ETL-Prozess eines DW als Spezialfall föderierter Datenbanken sehen. Für neue Datenquellen bedeutet der ETL-Prozess also manuellen Aufwand, der eine Interaktion mit einem Benutzer erfordert; im laufenden Prozess kann das Laden neuer Daten dann automatisch ausgeführt werden. Es stehen Tools zur Vereinfachung dieses Prozesses für die Anwender zur Verfügung, Beispiele dafür sind Talend[2] und IBM Data Stage[3].

4.3 Verwendung von Ontologien im ETL-Prozess

Die Idee, Ontologien zur Beschreibung von Objekten einzusetzen, ist weit verbreitet. Im DW-Bereich gibt es einen Vorschlag, Ontologien zu verwenden, um die Metadaten des Data Warehouses daraus abzuleiten [8]. In unserem Ansatz soll die Kopplung dieser beiden Gebiete auf andere Weise erfolgen: Aus den Klassifikationsattributen des DW soll eine Wissensbasis gebildet werden, die um Wörterbücher ergänzt wird.

[1] www.altova.com/mapforce.html

[2] www.talend.com

[3] www.ibm.com/software/data/infosphere/datastage

5. ZUSAMMENFASSUNG, AUSBLICK

Die flexible, durch situativ entstandenen Datenbedarf initiierte Integration bislang unerschlossener Datenquellen in ein Data Warehouse erfordert eine Anreicherung des ETL-Prozesses um interaktive Schritte. Um diesen Prozess für den Fachanwender handhabbar zu halten, bedarf es zusätzlicher Komponenten zur Speicherung und Nutzung von domänenspezifischem Wissen (*knowledge component*), die das Finden (*source identification*) und Integrieren (*data integration*) neuer Daten erleichtern bzw. erst ermöglichen.

Geleitet von Anwendungsszenarien wurde ein Konzept zur Architektur eines solchen Systems vorgestellt. Die Herausarbeitung technischer Herausforderungen zeigt den zu gehenden Weg: Die Details der einzelnen Komponenten sind zu konkretisieren, bislang nicht kombinierte Techniken zu verbinden und eine angemessene Nutzerschnittstelle zu entwickeln.

6. REFERENCES

[1] Baeza-Yates, R. und B. Ribeiro-Neto: *Modern information retrieval: the concepts and technology behind search*. Addison-Wesley, Pearson, Harlow [u.a.], 2. Aufl., 2011.

[2] Bauer, A. und H. Günzel: *Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung*. dpunkt-Verl., Heidelberg, 2. überarb. und aktualisierte Aufl., 2004. Literatur- und URL-Verz. S. 545–576.

[3] Courage, C. und K. Baxter: *Understanding Your Users: A Practical Guide to User Requirements Methods, Tools, and Techniques*. Morgan Kaufmann, 1. Aufl., 2005.

[4] Doan, A. und A. Y. Halevy: *Semantic Integration Research in the Database Community: A Brief Survey*. AI Magazine, 26(1):83–94, 2005.

[5] Inmon, W.: *Building the data warehouse*. Wiley, 2005.

[6] Kim, W. und J. Seo: *Classifying Schematic and Data Heterogeneity in Multidatabase Systems*. Computer, 24(12):12–18, Dez. 1991.

[7] Noy, N. F.: *Semantic integration: a survey of ontology-based approaches*. SIGMOD Rec., 33(4):65–70, Dez. 2004.

[8] Pardillo, J. und J.-N. Mazón: *Using Ontologies for the Design of Data Warehouses*. CoRR, abs/1106.0304, 2011.

[9] Rahm, E. und P. A. Bernstein: *A survey of approaches to automatic schema matching*. VLDB Journal, 10(4):334–350, 2001.

[10] Sheth, A. P. und J. A. Larson: *Federated database systems for managing distributed, heterogeneous, and autonomous databases*. ACM Comput. Surv., 22(3):183–236, Sep. 1990.

[11] Vitako: *IT-Monitor kommunal*. Vitako aktuell. Bundesarbeitsgemeinschaft der Kommunalen IT-Dienstleister e.V, 2007.