

Vergleich von Open-Source und kommerziellen Programmen zur Durchführung eines ETL-Prozesses



Exposé zur Diplomarbeit

Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät II
Institut für Informatik

Autor: Sascha Qualitz
Email:sascha.qualitz@informatik.hu-berlin.de
Betreuer: Prof. Dr. Ulf Leser

1 Motivation

1.1 Data Warehouse

Data-Warehouses¹ sind seit mehreren Jahrzehnten [Gl97] Werkzeuge zur Unterstützung der Entscheidungsfindung in Unternehmen. Sie integrieren Informationen aus verschiedenen meist heterogenen, operativen Informationssystemen, bereiten diese gemäß der Zielsetzung des Data-Warehouses auf und stellen sie Front-End-Systemen zur Darstellung für den Endnutzer zur Verfügung. Hierfür werden die extrahierten Daten mit Techniken aus dem Bereich Informationsintegration transformiert, um die für die Entscheidungsfindung und Analyse benötigte Datenqualität zu erhalten. Folgende Auswahl an Anforderungen an die Datenqualität, Detailtiefe und Struktur gehören nach [Fa11] dazu: die Eindeutigkeit der Daten, die Übernahme der Detaildaten samt entsprechender Metadaten, Vollständigkeit, Korrektheit, Konsistenz und Zuverlässigkeit. Diese Anforderungen an die Datenqualität müssen innerhalb des Extraktions-, Transformations- und Ladeprozesses, kurz ETL-Prozess, sichergestellt werden.

1.2 ETL-Prozess

Ein ETL-Prozess ist ein Vorgang, der aus mehreren unterschiedlichen heterogenen Informationssystemen Daten extrahiert, integriert und in ein Zielsystem lädt. Laut [Fa11] stellt dieser Vorgang innerhalb des Data-Warehouses den wichtigsten Teil beim Betrieb eines Data-Warehouses dar, findet aber auch in allen Szenarien, in denen Daten aus mehreren Quellen integriert und aufbereitet werden sollen Verwendung². Diese Daten werden im Verlauf in den Arbeitsbereich des DWs transportiert und dort für die Speicherung in der Basisdatenbank des DWs transformiert. In mehreren Schritten werden die Unterschiede in der Darstellung von Daten, wie zum Beispiel Datumsformaten oder Kodierungen bereinigt. Die extrahierten Daten werden validiert, gefiltert und unter Umständen mit zusätzlichen Informationen aufgewertet, um für eine konsistente und für das Data Warehouse relevante Datenbasis zu sorgen. Es werden dabei semantische und strukturelle Konflikte sowie Datenkonflikte aufgelöst, um ein einheitliches Schema aus den unterschiedlichen heterogenen Quellen zu generieren.

ETL-Prozesse können als gerichtete azyklische Graphen dargestellt werden. Naumann und Albrecht führen in [AlNa12] folgende übersetzte Definition ein.

¹ Data-Warehouse, kurz DW.

² In dieser Arbeit soll aber von einem Data-Warehouse als Zielsystem gesprochen werden.

Definition 1. *(ETL-Prozess) Ein ETL-Prozess besteht aus einer Menge von Transformationen T mit Eingangs- und Ausgangsschemata die verbunden miteinander einen DAG bilden. Sei $W = (V, E)$ so ein DAG der einen ETL-Prozess repräsentiert. V sei die Menge aller Knoten von W , welche die Transformationen repräsentieren. Die Kanten $e \in E \subseteq V \times V$ verbinden das Ausgangsschema einer Transformation mit dem Eingangsschema einer anderen Transformation, das heißt e repräsentiert ein geordnetes Paar von Transformationen.*

Am Ende des Transformationsprozesses stehen dann die integrierten Daten aller angebundenen Datenquellen in homogenen Schemata zur Verfügung. Der Abschluss dieses Vorgangs bildet das Laden der Daten in die Basisdatenbank und das Aktualisieren von Indizes und Data Marts beziehungsweise materialisierten Sichten [Va09].

2 Hintergrund

Für die meisten Firmen sind die Geschäftsdaten das höchste Gut. Diese aus den verschiedenen Geschäftsbereichen zu integrieren und zu analysieren, kann den Erfolg der Firma entscheidend beeinflussen. Der Aufwand für das Entwickeln von ETL-Prozessen zum Extrahieren, Transformieren und Laden der Daten in ein Data-Warehouse ist sehr hoch. ETL-Werkzeuge für das Befüllen von Data-Warehouses oder anderen Zielinformationssystemen verringern diesen Aufwand, in dem sie Schnittstellen- und Zielsystemheterogenitäten überwinden, Standardtransformationen wie zum Beispiel Datumsformatänderungen, Maßeinheitenumrechnungen oder Duplikaterkennung automatisiert vornehmen und graphische Benutzerschnittstellen bieten, um den ETL-Prozess mit so wenig Aufwand wie möglich zu spezifizieren. Auch Monitoringaufgaben können von ETL-Tools übernommen werden, um die durchgeführten Transformationen zurückverfolgen zu können. Im kommerziellen Bereich sind Produkte von Firmen wie Oracle, IBM, Microsoft, SAP und anderen zu erhalten, aber auch im Open-Source Bereich stehen einige ETL-Programme zur Verfügung, die vor allem für kleinere und mittlere Unternehmen von Interesse sind. Produkte aus diesem Bereich sind Pentaho Data Integration von der Firma Pentaho, Talend Open Studio von der Firma Talend, Jasper ETL von der Firma JasperSoft, sowie CloverETL von der Firma Javlin.

3 Ziel der Arbeit

Ziel der Arbeit ist es, eine Menge von ausgewählten Werkzeugen für die Durchführung von ETL-Prozessen zu untersuchen und zu bewerten. Es werden Open-Source Programme, wie zum Beispiel die in Kapitel 2 erwähnten Programme bezüglich ihres Funktionsumfangs

und ihrer Umsetzung ausgewählter, konkreter ETL Prozesse untersucht und miteinander verglichen. Mit diesen Ergebnissen wird ein kommerzielles Produkt zum Vergleich herangezogen, wobei dieses Produkt für den akademischen Betrieb kostenfrei zugänglich sein muss. Zur Evaluation wird auch gehören, den Ablauf eines ETL-Prozesses auf Basis eines Testsystems mit diesen Programmen durchzuführen, um praktische Probleme und Stärken zu ermitteln. Um die Programme richtig bewerten zu können, wird der ETL-Prozess in der Theorie beschrieben und die einzelnen Schritte erklärt werden.

4 Vorgehensweise

4.1 Definition des ETL-Prozesses

Der ETL-Prozess soll im Detail beschrieben und definiert werden, um die wichtigen Merkmale des Prozesses zu erkennen und den Funktionsumfang der untersuchten ETL-Werkzeuge realistisch einschätzen zu können.

4.2 Versuchsaufbau

Es werden die zu evaluierenden ETL-Programme gewählt und eruiert, welche Möglichkeiten es gibt, ein geeignetes Szenario als Testgrundlage zu finden, um die Performanz der Werkzeuge zu vergleichen. Eine mögliche Quelle hierfür ist das *“Transaction Processing Performance Council”*.

4.3 Vergleichsparameter

Für die Evaluation der Systeme auf Benutzerebene ist es nötig, Kriterien zu ermitteln nach denen die Programme verglichen und bewertet werden können. Mögliche Kriterien die herangezogen werden können, finden sich in [ThPe05] und [BaGu04]. Eine Auswahl dieser Kriterien, sind die Lizenz unter der die Programme vertrieben werden, die Werkzeuge für die Transformation der Daten und die Anzahl der Schnittstellen die das Programm anbietet, um unterschiedliche Quell- und Zielsysteme anzubinden.

4.4 Vergleich

Neben den in Abschnitt 4.3 beschriebenen Kriterien, sollen die Programme auch an, von der Benutzung, unabhängigen Parametern gemessen und bewertet werden. Diese Parameter werden im Detail auf Grundlage von [BaGu04] und [ThPe05] sowie wiederum dem *“Transaction Processing Performance Council”* und anderer maßgeblicher Literatur bestimmt werden. Dazu werden die Performanz des Testszenarios, d.h. zum Beispiel der

Durchsatz pro Datensatz, der Aufwand bei Änderung eines ETL-Szenarios, sowie die Verwaltung von Metadaten und ETL-Prozessen gehören. Die Ergebnisse werden zum Einen in die Beschreibung der einzelnen ETL-Werkzeuge und zum Anderen in eine tabellarische Gesamtübersicht über alle evaluierten Programme einfließen.

Literatur

- [AlNa12] ALBRECHT, A.; NAUMANN, F.: *Schema Decryption for Large Extract-Transform-Load Systems*. Conceptual Modeling, (116-125), 2012.
- [BaGu04] BAUER, A.; GÜNZEL, H.: *Data-Warehouse-Systeme : Architektur, Entwicklung, Anwendung*. dpunkt Verlag, 2. Aufl. Auflage, 2004.
- [Fa11] FARKISCH, K.: *Data-Warehouse-Systeme kompakt : Aufbau, Architektur, Grundfunktionen*. Springer, 2011.
- [Gl97] GLUCHOWSKI, P.: *Das aktuelle Schlagwort* Data Warehouse*. Informatik-Spektrum, (20):48–49, 1997.
- [ThPe05] THOMSEN, C.; BACH PEDERSEN, T.: *A Survey of Open Source Tools for Business Intelligence*. Data Warehousing and Knowledge Discovery, 3589:74–84, 2005.
- [Va09] VASSILIADIS, P.: *A Survey of Extract-transform - Load technology*. International Journal of Data Warehousing & Mining, 3(5):1–27, 2009.