Beitrag P: Sandra Schrauth, Radoslav Nedkov, Carsten Heidmann, Wassilios Kazakos, Andreas Abecker

Werkzeugunterstützung für ETL-Prozesse mit Geodaten

Sandra Schrauth, Radoslav Nedkov, Carsten Heidmann,
Wassilios Kazakos, Andreas Abecker

Disy Informationssysteme GmbH, Karlsruhe, vorname.nachname@disy.net

Abstract

Data Warehousing and Spatial Data Infrastructures (SDI) are becoming more and more accepted in public administrations, also in environment administrations and geo data authorities. Hence, the importance of professional ETL (extract - transform - load) processes for data acquisition, integration, cleansing, and storage is also growing. Though there are numerous ETL tools on the market since many years, not many of them provide comfortable functionalities for dealing with geo data. Hence Disy evaluated a couple of widespread Geo-ETL tools (Talend Open Studio, FME, GeoKettle, Oracle Data Integrator) with respect to their suitability for professional and sustainable ETL projects in eGovernment SDI contexts. It turned out that Talend Open Studio is in general very favorable, but still has weaknesses regarding geo data integration (Spatial ETL). So, Disy has developed a new Talend plug-in for Spatial ETL – which is presented in this paper.

Zusammenfassung

Data Warehousing und Geodateninfrastrukturen (GDI) verbreiten sich auch in öffentlichen Verwaltungen zunehmend. Dadurch steigt in diesem Bereich auch die Bedeutung sog. ETL-Prozesse für Datenimport, Datenintegration, Datenbereinigung und Datenspeicherung. Es sind bereits viele ETL-Werkzeuge seit vielen Jahren auf dem Markt, aber nur wenige haben auch komfortable Funktionen zum Umgang mit Geodaten. Da Geodaten in Umwelt-anwendungen aber häufig eine wichtige Rolle spielen, hat Disy einige weitverbreitete Geo-ETL Werkzeuge mit Blick auf ihre Eignung für professionelle und nachhaltige ETL-Projekte in öffentlichen GDI-Kontexten untersucht und verglichen – nämlich Oracle Data Integrator, GeoKettle, FME und Talend Open Studio. Dabei wird Talend Open Studio als insgesamt sehr empfehlenswertes Tool für unsere Anforderungen identifiziert, das aber noch deutliche Schwächen im Bereich Geodaten aufweist. Daher hat Disy ein neues Plug-In entwickelt, die Geospatial Integration für Talend.

1 Motivation und Überblick

Für Behörden und Unternehmen wird es immer wichtiger, die wachsende Menge an alphanumerischen Daten und Geodaten aus Fachanwendungen oder Sensoren für übergreifende Auswertungen, Datenportale und Berichtspflichten systematisch und möglichst automatisiert zu strukturieren und bereitzustellen.

Für die Realisierung von Datenintegrationslösungen in der öffentlichen Verwaltung in Deutschland setzt Disy seit einigen Jahren bei der Verarbeitung *alphanumerischer* Daten auf die Software Talend. Talend ist einer der Weltmarktführer im Bereich der ETL-Werkzeuge und hat sich auf die Integration großer Datenmengen spezialisiert.

In zahlreichen Projekten, gerade der Umweltverwaltung, spielen neben Sachdaten aber vor allem auch *Geodaten* eine entscheidende Rolle. Diese haben besondere Anforderungen, die bis dato in den meisten „klassischen“ ETL-Werkzeugen nur ansatzweise berücksichtigt sind. Dafür hat Disy gerade für die Geodatenverarbeitung in mehreren Projekten auch verschiedene andere Werkzeuge genutzt, wie z.B. insbesondere FME.

Um herauszufinden, ob es für Datenintegrationsaufgaben, die einen gleichermaßen guten Umgang mit alphanumerischen und mit Geodaten erfordern, ein klar zu präferierendes Werkzeug gibt, hat Disy zunächst die weiter verbreiteten Lösungen gesichtet und dann anhand eines praxisgetriebenen Kriterienkatalogs die Werkzeuge verschiedene bewertet. Nach einer ersten Auswahlrunde konnte man sich aufgrund der Randbedingungen für die effektive und professionelle Nutzung in unseren Kundenprojekten auf die Werkzeuge Talend Open Studio und FME fokussieren. In einer weiteren, tiefergehenden Untersuchung wurden diese beiden Werkzeuge genauer „unter die Lupe“ genommen. Es zeigte sich, dass (1) zwar FME die mächtigere, umfangreichere und komfortablere Geodatenverarbeitung besitzt, dafür aber (2) Talend als Gesamtlösung aus unserer Sicht für viele unserer großen und lang laufenden Kundenprojekte vermutlich die nachhaltigere Lösung darstellt. Da hier aber klare Nachteile gegenüber FME vorliegen, hat Disy eine Erweiterung von Talend realisiert, die Geospatial Integration for Talend.

Dieser Beitrag ist aufgebaut, wie folgt: In Kapitel 2 werden einige grundlegende Definitionen und Begriffe eingeführt. In Kapitel 3 wird die Vorgehensweise zur Identifikation eines geeigneten Werkzeugs vorgestellt. In Kapitel 4 werden die

Werkzeuge FME und Talend Open Studio eingehender untersucht und ein Zwischenfazit des Auswahlprozesses gezogen. Als Ergebnis wird die Realisierung der Geospatial Integration for Talend motiviert, welche in Kapitel 5 näher beschrieben wird. Kapitel 6 beendet den Beitrag mit Zusammenfassung und kleinem Ausblick.

2 Grundlagen

Eine sehr kurze, aber im Kern für das Verständnis völlig ausreichende Definition eines Data Warehouse formuliert [Rahm 2015] wie folgt:

Definition: Ein **Data Warehouse** (DW) ist eine für Analysezwecke optimierte zentrale Datenbank, die Daten aus mehreren, i.a. heterogenen Quellen zusammenführt und verdichtet (Integration und Transformation).

Verschiedene Autoren ergänzen noch diverse technische und zweckorientierte Merkmale (siehe z.B. [Inmon 1996; Bauer & Günzel 2013; Kimball & Ross 2013; Zeh 2003] und die gute Zusammenfassung bei [Wikipedia-1 2017]).

Wie sich schon aus der Bemerkung „Integration und Transformation“ als zentrale Aufgabe des DW ergibt, spielen die sog. ETL-Prozesse eine zentrale Rolle bei den Algorithmen für Aufbau und Betrieb eines DW. Wir folgen bei der Begriffsbildung hier [Hummeltenberg 2012]:

Definition: **ETL-Prozesse** umfassen das Extrahieren, Transformieren und Laden von Daten aus einem oder mehreren Quellsystemen in einen Zieldatenbestand inkl. Data Cleansing. ETL-Systeme bilden beim Data Warehousing die Datenschnittstelle zwischen operativen / externen Datenbeständen und Data Warehouse / Data Marts.

…

Bei einer materialisierten Datenextraktion, -integration und -aggregation wird zwischen den Phasen Extraktion, Transformation und Laden unterschieden und der Data Access und Integration Layer durch ETL-Systeme realisiert.

Während die Datenextraktion und das Laden zwar technisch anspruchsvoll sein können (insbesondere bei sehr großen Datenbeständen und Datenbeständen mit hoher Änderungsrate bzw. Datenströmen), finden sich jedoch die *konzeptionell* schwierigeren Aufgaben im Allgemeinen bei der Transformation. Hier führt [Hummeltenberg 2012] z.B. folgende Teilschritte an:

„…

1. *Auswahl der relevanten Daten, Elimination von Duplikaten*
2. *Schlüsselvergabe/-bereinigung*
3. *Überführung von CSV (Comma Separated Value)-Dateien in strukturierte Formate, XML/SQL-Konversion (XML, Structured Query Language)*
4. *Datenbereinigung, Integritätstests aufgrund Domänen oder vorgegebenen Mustern, Datenabgleich (Data Cleansing)*
5. *Überführung ereignisorientierter in periodenorientierte Größen, Währungsumrechnung, Aggregation, Kennzahlenermittlung u.a.*
6. *Datenintegration unterschiedlicher Quellen, Standardisierung, Datenergänzung (Datenfusion).*

…“ (vgl. z.B. auch [Wikipedia-2; Bauer & Günzel, 2013]).

Der Begriff des Geo Data Warehouse wird eher selten verwendet und kaum in der Fachliteratur einheitlich definiert. Wir erweitern daher den Begriff des Data Warehouse pragmatisch für unsere Zwecke, wie folgt:

Definition: Ein **Geo Data Warehouse** (GDW) oder Spatial Data Warehouse ist ein Data Warehouse, dessen Inhalte auch aus Daten mit Raumbezug bzw. Geodaten bestehen und das daher i.d.R. auch Operatoren und Optimierungen für räumliche Anfragen, Auswertungen und Analysen enthält.

Entsprechend beinhaltet die Menge der Quellsysteme eines GDW auch mindestens eine Quelle von Geodaten (GIS, Geodatenbank, Geodatendienst, Geodatendateien) und der Zieldatenbestand wird i.A. in einer Geodatenbank abgelegt.

Dabei verstehen wir unter einer **Geodatenbank**[32] eine Datenbank, die durch die Einbindung spezieller Datentypen, Datenstrukturen und Operatoren in der Lage ist, Geodaten effizient zu verwalten. Geodatenbanken verfügen vor allem über geeignete Sortier- und Suchverfahren, die eine effektive und schnelle Abfrage des

[32] Hier folgen wir [Martin et a. 2000]

Datenbestandes ermöglichen. Hierzu stellt sie für den Zugriff eine raumbezogene Abfragesprache bereit, die über räumliche Operatoren verfügt.

Beispiele für weitverbreitete Geodatenbanken zur Realisierung von GDW sind Oracle Spatial, PostGIS und SpatiaLite.

Da ein GDW in der Praxis häufig in eine komplexere Geodateninfrastruktur (GDI) eingebunden ist, sind oft auch noch weitere GIS- bzw. GDI-typische Software-Komponenten vorhanden, die in nicht geodaten-orientierten DW nicht vorkommen, wie insbesondere komplexere Metadatenbestände zu vorliegenden Daten oder dienste-basierte Schnittstellen für den Datenzugriff, die den Standards des OGC für den Geodatenaustausch entsprechen (WMS, WFS, …).

Beim Übergang vom Data Warehouse zum Data Warehouse mit räumlichen Daten ist notwendigerweise auch der Begriff der ETL-Prozesse zu erweitern:

Definition: **Spatial ETL-Prozesse** oder Geo-ETL Prozesse sind ETL-Prozesse, die auch Geodaten bzw. Daten mit Raumbezug verarbeiten können und typischerweise zum Realisieren eines Geo Data Warehouse genutzt werden.

Offensichtlich sind Spatial ETL-Prozesse also ETL-Prozesse mit folgenden Spezialisierungen bzw. Erweiterungen:

- *Extraktion*: kann Geodaten aus mindestens einem der Quellsysteme Geodaten-bank (wie PostGIS, Oracle Spatial / Locator, Esri personal geodb, MySQL spatial), GIS (wie ArcGIS Server, GE Smallworld), Geodatendienst (wie OGC WFS, SOS) oder Geodatendateien (wie Esri Shapefiles, GML, KML) einlesen.
- *Load*: kann Geodaten in mindestens eines der Zielsysteme Geodatenbank oder GIS schreiben.
- *Transformation*: kann mit gängigen Geodatenformaten bzw. Geodatentypen (wie Vektorgeometrien als WKT o.ä.) umgehen und umfasst Operatoren zur Verarbeitung von Geodaten bzw. zur räumlichen Datenverarbeitung. Dies umfasst beispielsweise:
    - Beachtung des verwendeten Koordinatenreferenzsystems (SRS) in Datenbeständen und Transformation zwischen verschiedenen SRS.
    - Operatoren für Geometrieobjekte, wie z.B. topologische Prädikate (liegt-in, beruehrt, ueberlappt, …), räumliche Verarbeitungen (Vereinigung,

Durchschnitt, Pufferung, …) oder auch fortgeschrittene Funktionen zur Weiterverarbeitung oder Datenqualitätssicherung (Geometriefehler finden und korrigieren, wie z.B. nicht geschlossene Formen).

Die weiter oben aufgeführten Teilschritte der Transformation in ETL-Prozessen tauchen genauso in Spatial ETL Szenarien auch auf, können aber in manchen Schritten anders, schwieriger oder aufwändiger zu berechnen sein, z.B.:

- Duplikatelimination oder Widerspruchserkennung kann mit Geodaten einfacher sein, wenn sich gewisse Sachdaten klar demselben Geometrieobjekt zuordnen lassen. Es kann aber auch wesentlich schwieriger sein, wenn in zwei Datensätzen leicht abweichende Geometrien auftauchen und zu entscheiden ist, ob damit in der Realwelt das gleiche Objekt gemeint ist.
- Aggregationen, Disaggregationen und Kennzahlenermittlung können aufwändiger sein, wenn verschiedene Datenreihen unterschiedliche räumliche Auflösungen verwenden oder sogar unterschiedliche räumliche Aggregations-hierarchien besitzen (z.B. politische Gliederungen wie Stadt-Kreis-Land vs. natrräumliche Gliederungen).
- Viele geometrische Verarbeitungsoperatoren besitzen hohe algorithmische Komplexität – verglichen mit Operatoren auf alphanumerischen Datentypen.
- u.v.m.

Data Warehouse und Geo Data Warehouse spielen in der öffentlichen Verwaltung und insbesondere in der Umweltverwaltung eine wachsende Rolle, vgl. z.B. [Albrecht & Bornhöft 2014; Hosenfeld & Albrecht 2015], mithin auch ETL und Spatial ETL Prozesse. Selbst ohne Aufbau eines persistenten DW/GDW sind die ETL-Funktionalitäten immer dan gefragt, wenn man (Geo-)Daten aus verschiedenen Quellen zusammenführt oder ineinander überführt. Deshalb war für Disy die Frage von strategischer Bedeutung, welches ETL-Werkzeug für unsere Kundenprojekte wohl am zukunftsfähigsten ist.

3 Auswahlprozess bei Disy für ein Geo ETL Werkzeug

Disy begegnet in seinen Kundenprojekten in den vergangenen Jahren zunehmend der Anforderung, für die strategische Werkzeugauswahl ein ETL Werkzeug zu nutzen bzw. zu empfehlen, das voraussichtlich auch langfristig in komplexeren Geo- und

Umweltdaten-Infrastrukturen z.B. großer Landes- oder Bundesbehörden nachhaltig und effizient verwendet werden kann.

Grundsätzlich fallen ja Import-, Integrations- und Transformationsaufgaben für größere Datenbestände oder Datenströme in verschiedenen Anwendungsfällen an, wie der Altdatenübernahme in neue Systeme, dem Zusammenführen von Datenbeständen, dem Aufbau von Auswertedatenbanken zur Effizienzsteigerung etc. Diese Aufgaben können einmalig, wiederholt oder regelmäßig auftreten. Je nachdem, wie komplex die auszuführenden ETL-Prozesse sind und wie häufig sie unter welchen Bedingungen vorkommen (z.B. Dynamik der Datenquellen und der Anwendungsfälle), sind spezielle ETL-Werkzeuge – im Gegensatz zu händisch ausprogrammierten Algorithmen – mehr oder weniger nützlich oder gar notwendig. Gerade wenn man sich in komplexen Software-/Daten-Umgebungen bewegt und auch längerfristig wiederholte ETL-Aufgaben zu erwarten sind, fällt die Auswahl eines optimalen ETL-Werkzeugs zunehmend ins Gewicht.

Daher hat Disy zunächst vier Werkzeuge als Kandidaten für die strategische Nutzung in Kundenprojekten identifiziert, die alle hervorragende Funktionalitäten vorweisen können, bei Disy-Kunden weithin zum Einsatz kommen und insgesamt eine große Bekanntheit und Nutzung aufweisen. Dies waren:

1) Oracle Data Integrator (ODI)[33]
2) GeoKettle[34]
3) Talend Open Studio[35]
4) FME[36]

ODI ist ein performantes und plattformunabhängiges, kommerzielles Werkzeug innerhalb des umfassenden Ökosystems von Oracle-basierten Produkten und Werkzeugen für Datenmanagement und -analyse, welches ausführliche Möglichkeiten zum Datenbankimport und für die Datenbankverwaltung anbietet. Geodatentypen werden erst ab Version 11g unterstützt. Die Transformationsmöglichkeiten für Geodaten sind

[33] Vgl. http://www.oracle.com/technetwork/middleware/data-integrator/overview/index.html
[34] Vgl. http://www.spatialytics.org/projects/geokettle/
[35] Vgl. https://de.talend.com/
[36] Vgl. https://www.safe.com/fme/key-capabilities/spatial-etl/

überschaubar bzw. nur über DB-Funktionalitäten lösbar. Das System erfordert eine gewisse Einarbeitung. Der Funktionsumfang ist für Geodaten nicht sehr hoch, kann aber durch eigene Knowledge Modules erweitert werden.

GeoKettle [Badard 2010; Badard et al. 2009] ist ein metadaten-gesteuertes ETL-Tool zur Verarbeitung von Geodatenbeständen[37], das auf dem weitverbreiteten Open Source BI-Stack Pentaho aufsetzt. GeoKettle besitzt einen guten und erweiterbaren Funktionsumfang, ist plattformunabhängig (Java-basiert) und nach unseren Erfahrungen intuitiv erlernbar und benutzerfreundlich. *Pentaho* und GeoKettle können allerdings nicht unabhängig voneinander aktualisiert werden und die Open Source Entwicklung führt zu unregelmäßigen, schwer vorhersehbaren Releases.

In einer ersten Bewertungsrunde wurden für beide Werkzeuge deutliche Stärken identifiziert. ODI ist natürlich naheliegend, wenn man sich innerhalb einer Oracle-dominierten Software-Landschaft bewegt und insbesondere, wenn man auch unabhängig von Geodatenverarbeitungen die entsprechenden Werkzeuge bereits intensiv nutzt und gut kennt. GeoKettle besticht durch seinen Funktionsumfang und seine Benutzerfreundlichkeit. Dennoch wurden beide Werkzeuge nicht zur intensiven weiteren Untersuchung ausgewählt. Während ODI sich in einer Umgebung, die kaum oder gar nicht Oracle-basiert ist, als proprietäres und komplexes Werkzeug kaum anbietet, hat GeoKettle zwar viele Vorteile, kann zurzeit aber nicht die vorhersagbaren regelmäßigen Release-Zyklen anbieten, die wir für unsere größeren Kunden mit regelmäßigen komplexen ETL-Aufgaben für notwendig halten.

Deshalb wurden nur FME und Talend Open Studio für die weitere Untersuchung herangezogen

- FME wegen der enorm hohen Verbreitung in der Geodatenwelt und wegen seines extrem großen Funktionsumfangs
- Talend wegen seines sehr starken Erfolgs in der Welt der Geschäftsdaten und der damit verbundenen enormen Dynamik seiner Entwicklung

[37] Siehe auch http://www.wheregroup.com/de/infobrief/01.2014/geokettle

Diese beiden Werkzeuge wurden anhand der in Abbildung 1 aufgeführten Bewertungskriterien näher untersucht. Die wesentlichen Ergebnisse werden im folgenden Kapitel dargestellt.

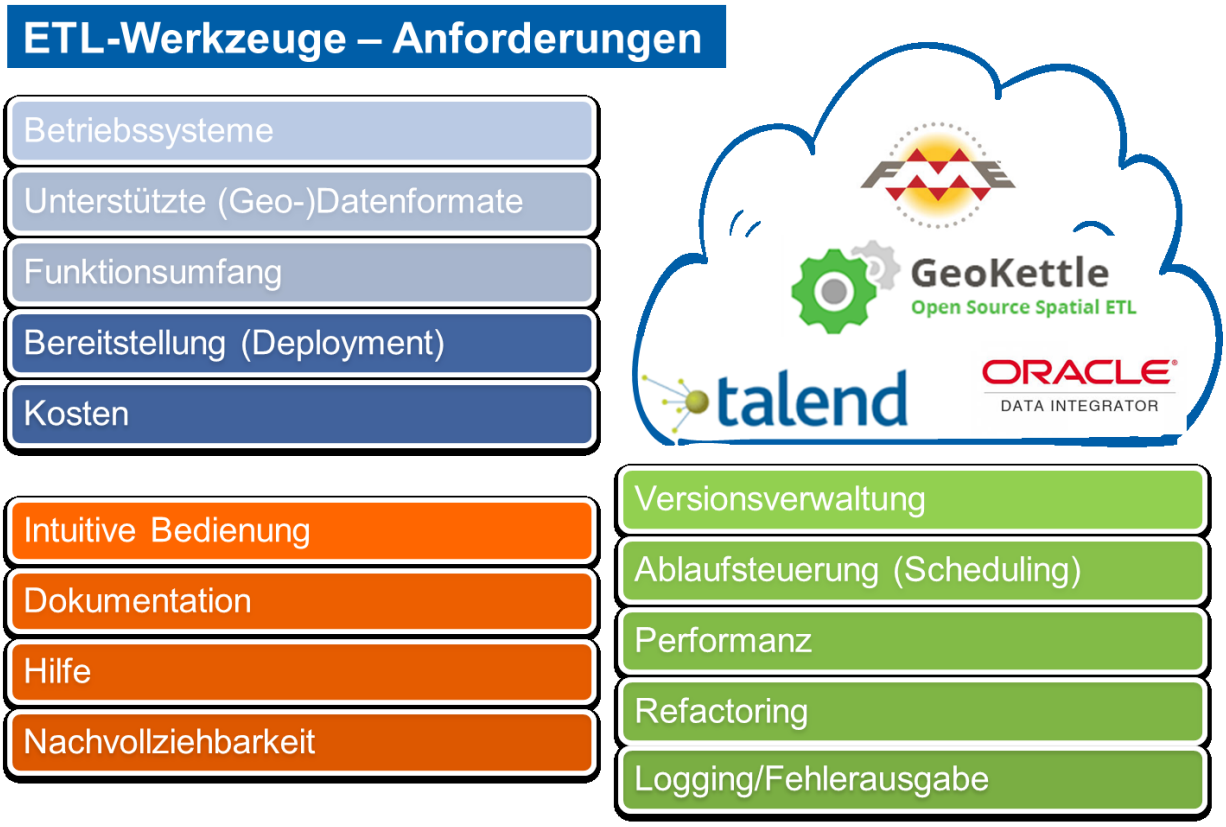

Abbildung 1: Bewertungskriterien für ETL-Werkzeuge

Die beiden ETL-Werkzeuge FME und Talend wurden anhand der oben aufgelisteten Kriterien eingehender untersucht, um ihr jeweiliges strategisches Potenzial als ETL-Werkzeug in komplexen, professionellen Datenintegrationsaufgaben für Szenarien mit umfangreichen Sach- und Geodaten abzuschätzen. Dabei kann Disy als Mittelständler nicht leisten, mit sehr hohem Zeitaufwand einen Toolvergleich zu erstellen, der jedem akademischen Qualitätsanspruch gehorcht. Gleichwohl können Mitarbeiter, die seit Jahren mit GDW und Spatial ETL in praktischen Kundenprojekten arbeiten, versuchen, eine möglichst faire, objektive und praxisorientierte Einschätzung zu liefern, die den aktuellen Sachstand im Licht der spezifischen Anforderungen des Einsatzes bei Disy reflektiert. Insbesondere hatten Disy-Mitarbeiter mit beiden Werkzeugen bereits im Zuge von Kundenprojekten Praxiserfahrungen gesammelt:

- FME wurde beispielsweise genutzt, um ETL-Prozesse zum Aufbau der Landesdatenbank Wasser des Niedersächsischen Landesbetrieb für Wasserwirtschaft,

Küsten- und Naturschutz (NLWKN) zu realisieren.[38] FME kam auch beim Aufbau der kommunalen GDI für die Stadt Baden-Baden zum Einsatz.[39]

- Talend Open Studio wurde für die Realisierung eines Data Warehouse zur Unterstützung des Portals „Artdaten Online“ des Sächsischen Landesamts für Umwelt, Landwirtschaft und Geologie (LfULG)[40] genutzt, beim Datenbank-Redesign für den „Energieatlas Bayern“[41] und für die Datenintegration von Unternehmensdaten für den Deutschen Industrie- und Handelskammertag (DIHK e.V.).[42]

4 Tiefergegende Analyse von FME und Talend

4.1 Das ETL-Werkzeug FME

FME Desktop [con terra 2015] ist ein Produkt der kanadischen Firma Safe Software Inc. und wahrscheinlich zurzeit das weltweit meistgenutzte Spatial ETL-Werkzeug zur Integration, Bearbeitung und Qualitätssicherung räumlicher Daten. Unterschiedlichste räumliche Datenquellen lassen sich schnell und effizient in einen FME Prozess importieren, umstrukturieren und in ein benutzerspezifisches Zieldatenmodell überführen. ETL-Prozesse werden mit FME Desktop erstellt und können dann über FME Server oder FME Cloud anderen Anwendern als Dienste zur Verfügung gestellt werden.

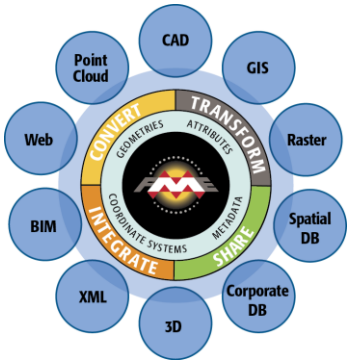

Abbildung 2: FME im Überblick (Quelle: safe.com)

[38] Vgl. https://www.disy.net/aktuelles/newsletter/newsletterartikel/artikel/2894.html
[39] Vgl. https://gispoint.de/news-einzelansicht/1887-disy-buerger-gis-fuer-baden-baden.html
[40] Vgl. https://www.disy.net/nc/aktuelles/newsartikel/artikel/3044.html
[41] Vgl. https://www.disy.net/nc/aktuelles/newsartikel/artikel/3023.html
[42] Vgl. https://www.disy.net/nc/aktuelles/newsartikel/artikel/2972.html

FME Desktop wird in verschiedenen Lizenzstufen angeboten, je nachdem, welche Datentypen verarbeitet werden sollen (z.B. Daten aus Esri GIS-Produkten, aus GE Smallworld oder aus Geodatenbanken).

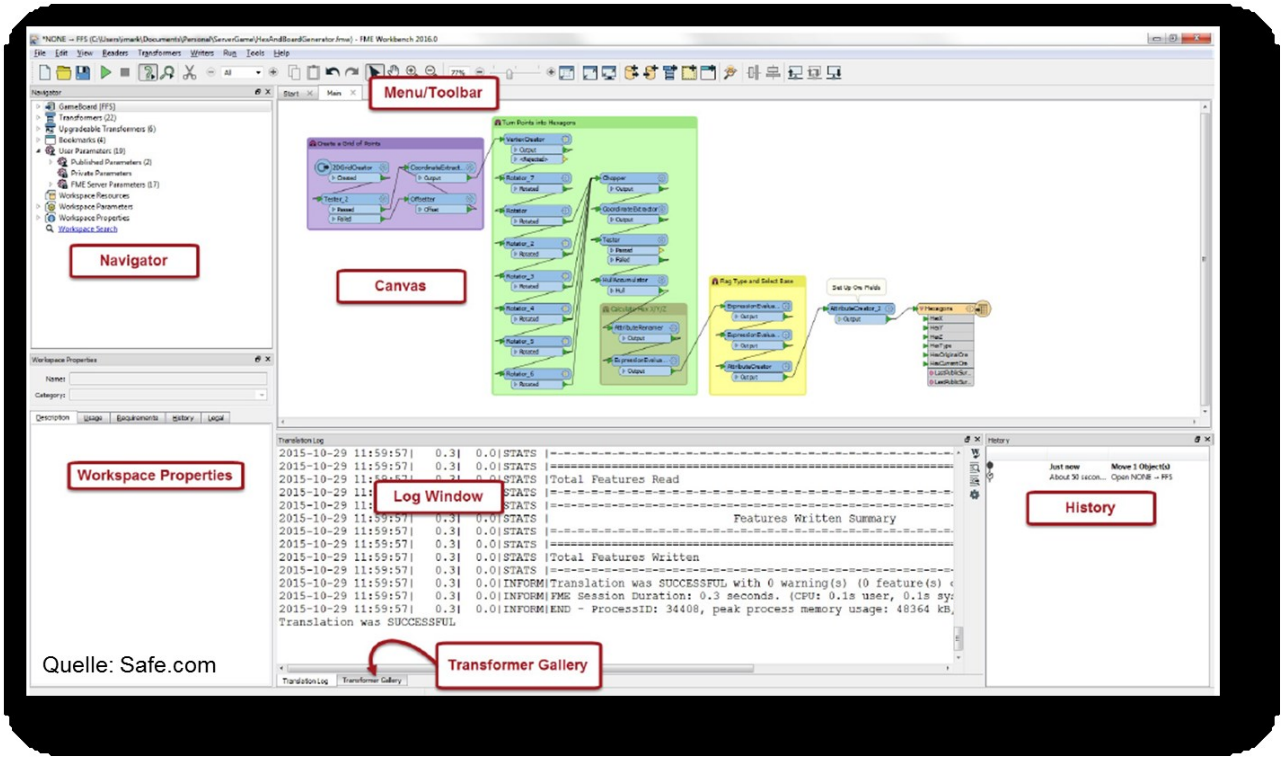

Abbildung 3: Die FME Workbench (Quelle: safe.com)

FME besteht aus verschiedenen Komponenten:

- *Quick Translator* erlaubt die einfache Umsetzung von Standardkonvertierungen
- *Data Inspector* ist ein Datenviewer für Geodaten
- *FME Workbench* ist ein graphisches Modellierungswerkzeug für ETL-Prozesse
- *Objects API* und *Plug-in Developer Kit* ermöglichen die Integration von FME-Funktionalitäten in eigene Anwendungen oder das Hinzufügen eigener Datenformate

Wir werden verschiedene Eigenschaften von FME weiter unten beim Vergleich mit Talend noch vorstellen. Vorneweg aber einige der augenfälligsten Merkmale:

- **Datenformate**: FME unterstützt eine enorme Anzahl von Datenformaten, mit einem deutlichen Fokus auf räumlichen und Geodaten; insgesamt über 300 Datenformate, darunter CAD-Formate, GIS-Formate, Geodatenbanken, Rasterdaten,

OGC-Webdienste, 3D-Daten oder BIM-Daten. Dagegen liegen Sachdaten nicht im Fokus der Betrachtung.

- **Verarbeitungsfunktionen**: ebenso werden mehr als 400 vordefinierte *Transformer* zur Bearbeitung raumbezogener Informationen angeboten.
- **Deployment**: FME erlaubt die Erstellung von Batchfiles; zur Abarbeitung ist aber eine FME-Lizenz (FME Server, FME Cloud) notwendig.
- **Logging**: es wird ein Logfile erzeugt.

4.2 Das ETL-Werkzeug Talend Open Studio

Die in den USA angesiedelte Firma Talend bietet ein komplexes Ökosystem von teilweise kostenlosen und teilweise kommerziellen Datenintegrationsprodukten an, mit Schwerpunkten wie z.B. Datenqualität, Big Data oder Master Data Management. Zentral und grundlegend ist die kostenlose Open Source Lösung Talend Open Studio for Data Integration.

Talend Open Studio ist ein Werkzeug für grafisches Design und Verwaltung von Datenintegrationsabläufen (Jobs) und Geschäftsmodellen. Talend Open Studio basiert auf Eclipse IDE und arbeitet als Codegenerator mit der Codeausgabe in Java. Dadurch ist es einfach möglich, in Talend modellierte Datenintegrations-Jobs auf jedem System mit Java-Laufzeitumgebung (JRE) auszuführen.

Datenintegrationsabläufe werden grafisch repräsentiert und modelliert und setzen sich aus Komponenten zusammen. Talend Jobs können in anderen Jobs als Subjobs ausgeführt werden.

Ein zentrales Konzept in Talend ist die zentrale Verwaltung von Informationen zu Datenverbindungen und Parametern wie Pfaden, DB-Verbindungsoptionen u.ä.. Solche Metadaten sind bequem projektübergreifend synchronisierbar.

Alle Parameter einer Datenbankverbindung können per Knopfdruck in sog. „Kontexte“ umgewandelt werden. Dadurch wird es sehr einfach, bei der Jobausführung zwischen Entwicklungs-, Test- und Produktivumgebung umzuschalten.

Datenschemata (Attribute und Datentypen) werden zentral definiert. Änderungen können automatisch in alle Jobs übernommen werden, in denen die Daten verwendet werden (einfaches Refactoring).

Diese Funktionen zeigen schon, dass Talend sehr komfortabel für den Umgang in großen professionellen IT-Umgebungen und mit komplexen ETL-Projekten geeignet ist. Diese Vorteile z.B. beim Metadaten-Management, Release Management, Deployment und Refactoring kommen beim Übergang zur kommerziellen Version Talend Enterprise Data Integration noch besser zum Tragen: Unterstützung der Versionsverwaltung (SVN, GIT), Datenvorschau, Distance Run, Jobvergleich und die Erstellung von Vergleichsdateien zum Testen werden dann bspw. zusätzlich angeboten.

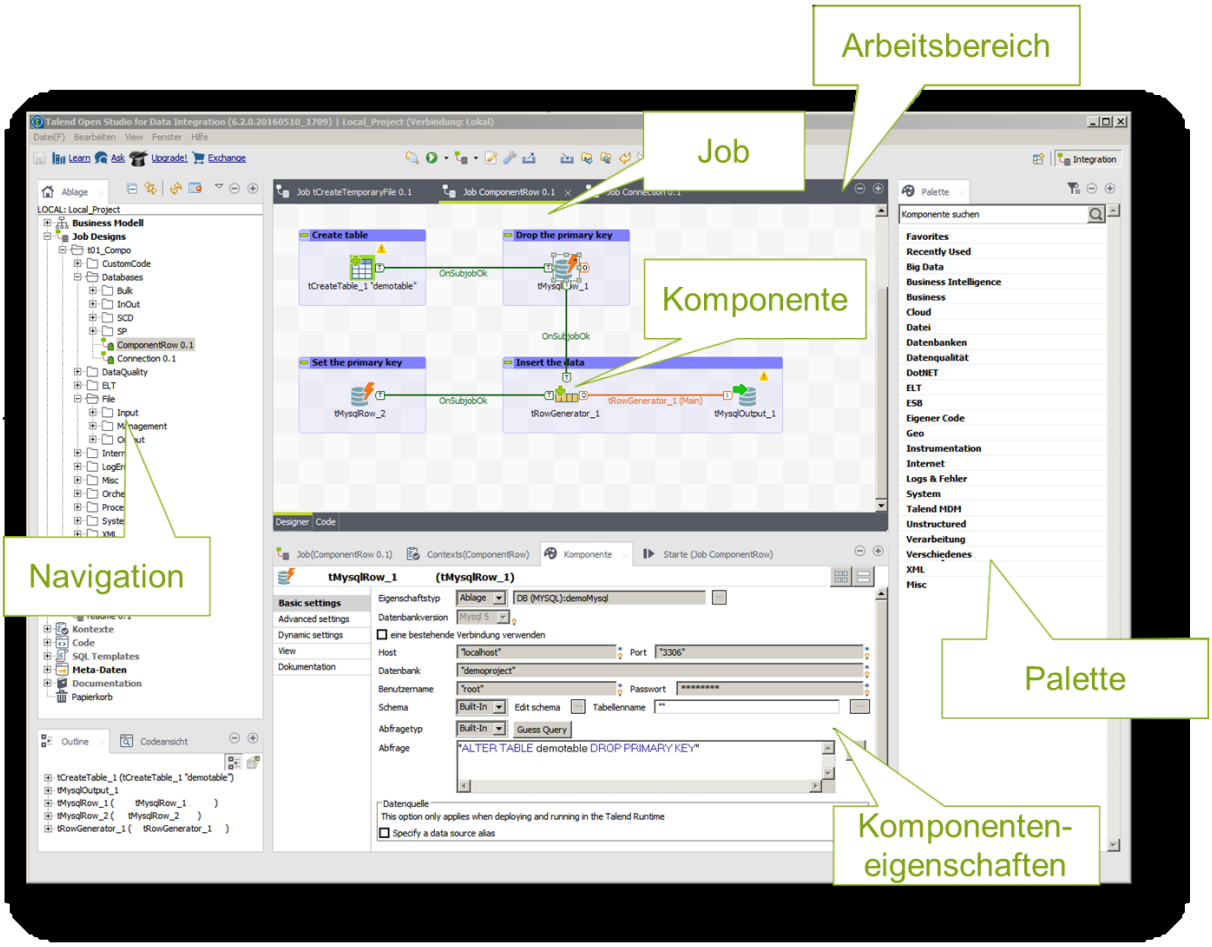

Abbildung 4: Talend Open Studio

Mit Bezug auf die oben bereits für FME angeführten Merkmale lässt sich sagen:

- **Datenformate**: Wie FME unterstützt Talend die gängigsten DB- und Datei-Formate, wobei Talend den Schwerpunkt auf die Sachdaten legt. Geodaten werden von Haus aus nicht unterstützt; Talend kann jedoch durch GeoSQL und durch die Open Source Spatial Extension von CamptoCamp erweitert werden.

- **Verarbeitungsfunktionen**: es werden über 450 Komponenten vorgefertigt mitgeliefert, jedoch praktisch ausschließlich für Sachdaten; beliebige Erweiterungen durch Java-Code sind möglich.
- **Deployment**: Jobs werden als Java Build erzeugt und können als Standalone oder Webservice (Axis) mit einer JRE ausgeführt werden.
- **Logging**: Talend hat zahlreiche Möglichkeiten für Logging und Fehlerausgabe.

4.3 Zwischenfazit

Auf der Basis umfangreicher Untersuchungen und Bewertungen der beiden näher betrachteten Werkzeuge kam das in Abbildung 5 zusammengefasste Ergebnis zustande – die rot hervorgehobenen Zeilen sind für Disy von besonderer Bedeutung.

FME und Talend sind beides hervorragende ETL-Werkzeuge mit sehr großem Funktionsumfang, sehr guter Erlernbarkeit und Benutzer-Unterstützung sowie hohem professionellem Qualitätsstandard.

FME bleibt weiterhin das Werkzeug der Wahl, wenn man sich fast ausschließlich im Bereich der räumlichen Geodaten bewegt und dort umfangreiche oder komplexe ETL-Prozesse durchführen muss, insbesondere im Zusammenhang mit GIS-spezifischen Phänomenen und Tools wie z.B. BIM, 3D-Modellen, Rasterdaten o.ä.

In vielen Kundenprojekten von Disy spielen umfangreiche Sachdaten aber eine gleichberechtigte Rolle neben Geodaten. In diesem Bereich ist Talend hingegen fast „unschlagbar“. Hinzu kommen signifikante Vorteile von Talend im unteren Bereich der Übersichtstabelle in den Dimensionen Versionsverwaltung, Wiederverwendbarkeit, Refactoring und Logging. Hier spielt Talend seine großen Stärken aus. Als Werkzeug für die schnelle und einfache Definition – auch großer – ETL-Projekte und deren Nutzung in komplexen Software-Umgebungen bietet Talend hervorragende Möglichkeiten zur Produktivitätssteigerung und Qualitätssicherung. Aus Sicht der professionellen Software-Entwicklung wird hier ein sehr hohes Niveau erreicht. Da sich Talend in seinen kommerziellen Produkten auch gerade mit modernsten Ansätzen aus Cloud-Computingund Big Data befasst, können auch im Bereich „Zukunftssicherheit“ Punkte gesammelt werden.

Allerdings ist die Dimension Geodaten bei Talend – auch erweitert mit der CamptoCamp Lösung – noch deutlich ausbaufähig. Das Werkzeug ist hier aus Sicht

von Disy für Anwender mit professionellen GIS-Ansprüchen noch nicht konkurrenzfähig, weder im Umfang noch in der Umsetzungsqualität. Deshalb hat Disy die im folgenden Kapitel skizzierte Lösung Geospatial Integration for Talend entwickelt.

	Talend	FME
Betriebssysteme	++	o
Datenformate	++	+
Geodatenverarbeitung	o	++
Funktionsumfang	++	+
Deployment	++	+
Bedienung	++	++
Dokumentation	+	++
Hilfe	+	++
Nachvollziehbarkeit	+	+
Versionsverwaltung	++	-
Wiederverwendbarkeit	++	o
Scheduling	++	+
Performanz	++	+
Refactoring	++	-
Logging	++	-

Abbildung 5: Zusammenfassende Bewertung aus unseren Untersuchungen

5 Disy's Geospatial Integration für Talend

Da die Talend Erweiterung für Geodaten aus unserer Sicht viele Wünsche offen ließ, wurde in Kundenprojekten zunächst für die Verarbeitung von Geodaten auf zusätzliche Werkzeuge zurückgegriffen. Daraus entstand der Wunsch nach einem Tool mit mächtigen Spatial ETL Funktionalitäten, das sich so nahtlos wie möglich in den bewährten Talend-Prozess einbinden lässt, so dass eine einheitliche Arbeitsweise für alle Daten angewendet werden kann.

Abbildung 6: Disy's Geospatial Erweiterung von Talend in der Werkzeugleiste

Deshalb hat Disy das Plug-in GeoSpatial Integration für Talend entwickelt, das im Zusammenspiel mit der bereits existierenden Talend-Software Daten vom Typ „Geometrie“ erkennt und für diese zusätzliche Kalkulatoren und räumliche Operatoren bereitstellt. Dadurch können alphanumerische Daten geometrisch angereichert und Geodaten einfach in Datenintegrationsprozesse eingebunden werden.

Das neue Plug-in wird in die Talend-Umgebung direkt eingebunden und erweitert somit die vorhandene Werkzeugleiste nahtlos. Der Benutzer sieht die zusätzlichen Datenquellen sowie die neuen Operatoren, die er per Drag-and-drop in das Arbeitsfenster übernehmen kann. Abhängig von der aktuell genutzten Komponente kann er weitere Einstellungen vornehmen oder zusätzliche Berechnungen durchführen.

Weit verbreitete relationale Datenbanken wie Oracle oder PostgreSQL unterstützen bereits seit einigen Jahren mit Oracle Locator/Spatial oder PostGIS räumliche Datentypen und Operatoren für die Verarbeitung von Geodaten. Mit dem von Disy entwickelten Plug-in GeoSpatial Integration für Talend können nun diese Geodaten direkt in Talend Datenintegrationsprozessen mit eingebunden werden. Konkret unterstützt das Plug-in aktuell folgende Datenbanken und Formate: Oracle Locator und Spatial, PostgreSQL mit PostGIS, SQLite mit SpatiaLite sowie Shapefiles und WKT (Well-Known-Text). Weitere Konnektoren für SAP HANA oder ArcGIS Server sind geplant.

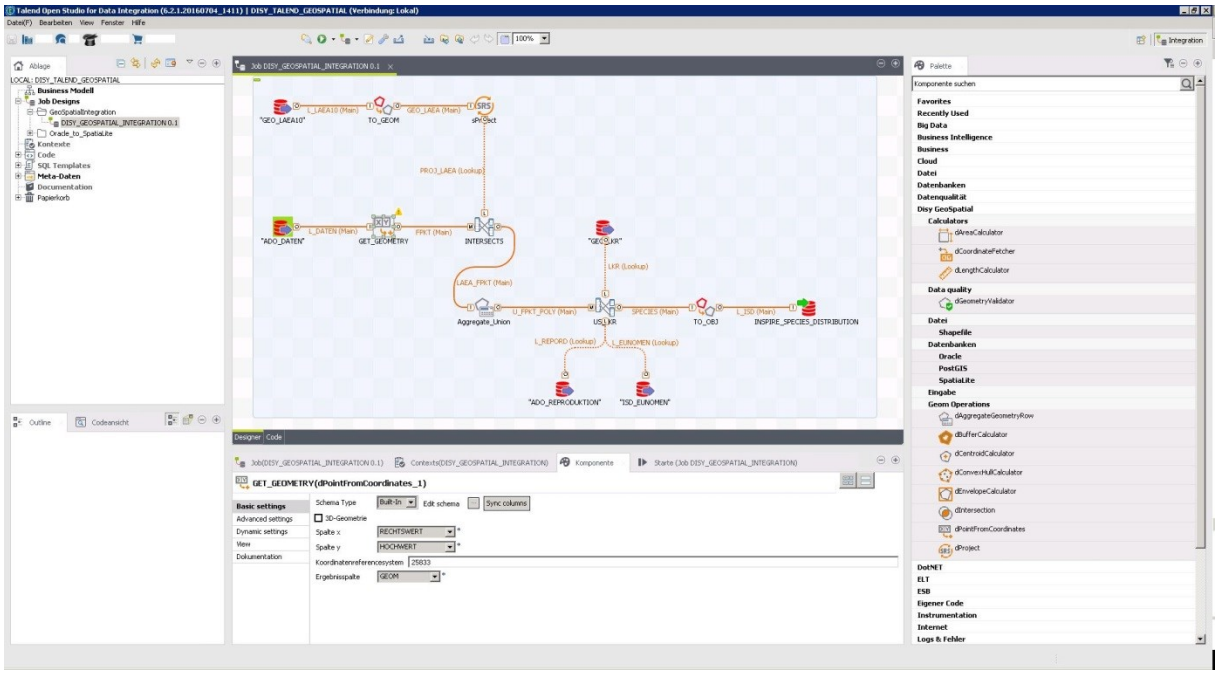

Abbildung 7: Beispiel Screenshot eines Spatial ETL Prozesses in Talend unter Nutzung des Disy Plug-ins

Hinzu kommt eine Vielzahl an Komponenten und räumlichen Operatoren, mit deren Hilfe Geooperationen durchgeführt werden. Hierzu zählen zurzeit:

- Längen- und Flächenberechnungen,
- Umwandlung von X-, Y- und Z-Koordinaten in 2D/3D-Punktgeometrien,
- Berechnung von Centroiden,
- Pufferung von Punkten, Linien und Flächen,
- Verschneidung von Geometrien,
- Berechnung einer Bounding Box (envelope) oder einer konvexen Hülle einer oder mehrerer Geometrien,
- Verbindung von Punkten zu Linien bzw. von Linien zu Flächen,
- Transformation der Koordinaten zwischen unterschiedlichen Koordinaten-systemen,
- algorithmische Vereinfachung von komplexen Geometrien
- Validierung von Eingangsdaten (z. B. Shapefiles).

Die Geooperationen stehen direkt als Talend-Routinen und/oder -Komponenten zur Verfügung. Der Funktionsumfang ist jederzeit erweiterbar.

Für Talend Open Studio wird die Nutzung von GeoSpatial Integration kostenlos zur Verfügung gestellt. Unternehmen und Behörden, die die Lösung in Produktivsystemen oder zusammen mit Talend Data Integration oder der Talend Data Management Platform einsetzen möchten, wird ein jährliches Abonnement (Subscription) für professionellen Support und Zusatzfunktionen zur Datenqualität, Visualisierung und mehr angeboten.

6 Zusammenfassung und Ausblick

In unserer Arbeit für Kunden der öffentlichen Verwaltung werden in den vergangenen Jahren Prozesse zum Spatial ETL zunehmend bedeutsamer, auch und gerade im Bereich der Umweltinformatik bzw. UIS.Um herauszufinden, ob es für Datenintegrationsaufgaben, die einen gleichermaßen guten Umgang mit alphanumerischen und mit Geodaten erfordern, ein klar zu präferierendes Werkzeug gibt, hat Disy zunächst die weiter verbreiteten Lösungen gesichtet und dann anhand eines praxisgetriebenen Kriterienkatalogs die Werkzeuge Talend, FME, GeoKettle[43] und Oracle Data Integrator[44] bewertet. In einer ersten Auswahlrunde konnte man sich aufgrund der Randbedingungen für die effektive und professionelle Nutzung in unseren Kundenprojekten auf die Werkzeuge Talend und FME fokussieren. In einer weiteren, tiefergehenden Untersuchung wurden diese beiden Werkzeuge genauer „unter die Lupe“ genommen. Es zeigte sich, dass

- zwar FME die mächtigere, umfangreichere und komfortablere Geodatenverarbeitung besitzt,
- dafür aber Talend als Gesamtlösung (mit der Stärke allerdings bei alphanumerischen Daten) aufgrund von Software-Engineering Stärken für viele unserer großen und lang laufenden Kundenprojekte vermutlich die nachhaltigere Lösung darstellt.

Allerdings bewerteten wir die existierende Spatial Lösung für Talend als ausbaufähig. Disy hat deshalb mit dem „Disy GeoSpatial Integration for Talend“ eine mächtige neue Lösung geschaffen.

---

[43] http://www.spatialytics.org/projects/geokettle/
[44] http://www.oracle.com/technetwork/middleware/data-integrator/overview/index.html

Für den Aufbau von Data Warehouses oder Auswertedatenbanken mit Geodaten ergeben sich durch diese Lösung zwei zentrale Vorteile: (1) Alle benötigten Datenarten können ohne Technologiebruch mit einem statt wie bisher mit mehreren Werkzeugen verarbeitet werden. Dies spart organisatorischen Aufwand zur Zusammenführung der Werkzeuge, reduziert den Einarbeitungsaufwand und stellt ein konsistentes Vorgehen bei alphanumerischen Daten und Geodaten sicher. (2) Bewährte und praxiserprobte ETL-Technologien, wie sie von Talend bereits für Sachdaten angeboten werden, können nun auch für die Geodatenverarbeitung genutzt werden. Neben der sehr umfassenden Menge an Datenquellen, Komponenten und Routinen, die mit GeoSpatial Integration mitgeliefert werden, gehören hierzu vor allem auch Funktionen, die Talend bereits mitbringt. Besonders hervorzuheben sind z.B. Funktionen zur Versionsverwaltung, zum Metadatenmanagement, zum Arbeiten in verteilten Teams und Releasemanagement, zum Refactoring sowie zur zentralen Administration, dem Load-Balancing oder sogar der Big-Data-Verarbeitung.

Die dargestellten Arbeiten wurden mit Unterstützung des FuE-Projekts WIRE durchgeführt. In diesem Rahmen sollen noch weitere Möglichkeiten untersucht werden, um mit Methoden des Semantic Web und des Maschinellen Lernens intelligente Werkzeuge zur (teil-)automatisierten, lernenden Geodatenintegration und Qualitätsverbesserung von Geodaten zu schaffen. Weitere Aufgabenfelder, die im Rahmen des Projekts betrachtet werden sollen, sind zum Beispiel:

- Lernende Verfahren zur Unterstützung des Datenschema-Matchings
- Automatische Identifikation des in einem Geodatenbestand verwendeten Koordinatenreferenzsystems (SRID)
- Lernende Verfahren zum Auffüllen von Datenlücken
- Bessere Methoden zum Geocoding

Insgesamt ergeben sich also spannende Perspektiven, um einerseits die Praxistauglichkeit und den operativen Nutzen „einfacher“ Spatial ETL-Ansätze weiter zu untersuchen und andererseits noch innovativere Lösungsansätze auf ihre Machbarkeit hin abzuklopfen.

Danksagung: Die Arbeiten an innovativen Methoden und Werkzeugen für Geo-ETL-Prozesse werden vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des KMU-innovativ Projekts WIRE (Intelligentes Werkzeug für Qualitäts-verbesserungen im multi-dimensionalen Datenwürfel, FKZ 01IS16039) unterstützt. WIRE wird von Disy koordiniert und zusammen mit dem FZI Forschungszentrum Informatik am Karlsruher Institut für Technologie bearbeitet.

7 Literaturverzeichnis

Albrecht, M.; Bornhöft, D. (2014): Mit Strategie zu neuen Architekturen – Cadenza als strategische Kernkomponente in der IT des Ministeriums für Energiewende, Landwirtschaft, Umwelt und ländliche Räume des Landes Schleswig-Holstein. In: K. Weissenbach, W. Schillinger, R. Weidemann (Hrsg.): *F+E-Vorhaben MAF-UIS / Moderne anwendungsorientierte Forschung und Entwicklung für Umweltinformationssysteme, Phase II 2012/2014.* Karlsruhe: KIT. KIT SCIENTIFIC REPORTS No. 7665, S. 115 – 125.

Badard, T. (2010): GeoKettle: A powerful open source spatial ETL tool. In: *FOSS4G 2010, Barcelona / Spain.* URL: https://de.slideshare.net/tbadard/geokettle-a-powerful-open-source-spatial-etl-tool-5193932 . Letzter Zugriff: 06.08.2017.

Badard, T.; Dubé, E.; Diallo, B.; Mathieu, J.; Ouattara, M. (2009): GeoKettle: A powerful open source spatial ETL tool. In: *FOSS4G 2009, Sydney / Australia.* URL: https://de.slideshare.net/tbadard/geokettle-a-powerful-open-source-spatial-etl-tool . Letzter Zugriff: 06.08.2017.

Bauer, A.; Günzel, H. (2013): *Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung, 4. Auflage.* dpunkt, 2013, ISBN 3-89864-785-4.

Con terra GmbH (Hrsg.) (2015): *FME Desktop - Das deutschsprachige Handbuch für Einsteiger und Anwender.* Berlin, Offenbach: Wichmann Verlag. ISBN 978-3-87907-591-1.

Hosenfeld, F.; Albrecht, M. (2015): Energy Atlas Schleswig-Holstein. In: *Adjunct Proceedings of the 29th EnviroInfo and 3rd ICT4S Conference, Copenhagen / Denmark*. URL: http://enviroinfo.eu/sites/default/files/pdfs/vol9073/0122.pdf . Letzter Zugriff: 06.08.2017.

Hummeltenberg, W. (2012): ETL. In N. Gronau et al. (Hrsg.): *GITO Online Lexikon Enzyklopädie der Wirtschaftsinformatik.* URL: http://www.enzyklopaedie-der-wirtschaftsinformatik.de/lexikon/daten-wissen/Business-Intelligence/ETL . Letzter Zugriff: 02.08.2017.

Inmon, W.H. (1996): *Building the Data Warehouse.* John Wiley & Sons, 1996, ISBN 978-0-471-14161-7.

Kimball, R.; Ross, M. (2013): *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd Edition*. Wiley.

Martin, C.; Bischof, N.; Eiblmaier, M. (Hrsg.) (2000): Geodatenbank. In: *Online Lexikon der Geowissenschaften.* Heidelberg: Spektrum Akademischer Verlag. URL: http://www.spektrum.de/lexikon/geowissenschaften/geodatenbank/5586 . Letzter Zugriff: 06.08.2017.

Rahm, E. (2015): *Data Warehouses. Einführung.* Vorlesungsskript, Universität Leipzig. URL: dbs.uni-leipzig.de/file/dw-kap1.pdf . Letzter Zugriff: 24.07.2017.

Wikipedia-1: *Data Warehouse.* URL: https://de.wikipedia.org/wiki/Data-Warehouse . Letzter Zugriff: 24.07.2017.

Wikipedia-2: *ETL-Prozess.* URL: https://de.wikipedia.org/wiki/ETL-Prozess . Letzter Zugriff: 02.08.2017.

Zeh, T. (2003): Data Warehousing als Organisationskonzept des Datenmanagements. Eine kritische Betrachtung der Data-Warehouse-Definition von Inmon. In: *Informatik – Forschung und Entwicklung.* 18, Nr. 1, 2003. URL: http://tzeh.de/abstract_dw.htm Letzter Zugriff: 06.08.2017