Geodatenmanagement und -harmonisierung mit Open-Source-Tools

Jens SCHAEFERMEYER
WhereGroup GmbH & Co. KG, Bonn · jens.schaefermeyer@wheregroup.com

Zusammenfassung

Die Harmonisierung von Geodatenbeständen ist nicht erst seit INSPIRE ein großes Thema bei Geodaten haltenden oder -produzierenden Organisationen. Immer wieder müssen Geodatenbestände zwischen verschiedenen Formaten oder Schemata transformiert werden. Bisher wurden diese Aufgaben i. d. R. mit kostenpflichtigen ETL-Tools (Extract, Transform, Load) bearbeitet. Seit einiger Zeit gibt es für diese Zwecke auch freie Softwarepakete. Im vorliegenden Vortrag soll die Arbeitsweise mit einem dieser Tools exemplarisch aufgezeigt werden.

GeoKettle ist ein ETL-Programm für räumliche Daten. GeoKettle basiert auf der Open-Source-Software *Pentaho Data Integration (Kettle)* und ist mit der LPGL lizenziert. GeoKettle unterstützt dabei u. a. die Open-Source-Bibliotheken GeoTools, deegree und gdal/ogr sowie sextante.

1 GeoKettle

GeoKettle ist ein ETL-Programm für räumliche Daten. ETL steht für **Extract**, **Load** und **Transform**. GeoKettle basiert auf der Open-Source-Software *Pentaho Data Integration (Kettle)* und ist mit der LPGL lizenziert. GeoKettle unterstützt dabei u. a. die Open-Source-Bibliotheken GeoTools, deegree und gdal/ogr und sextante.

Die Modellierung von Datenoperationen wird in GeoKettle in Transformationen und Jobs vorgenommen. In den Transformationen werden Daten verarbeitet. Daher hat eine Transformation immer einen Dateninput, der entweder aus einer originären Datenquelle oder aus der Übergabe von Daten aus einer vorherigen Transformation bzw. aus einem vorherigen Job besteht. Die Abbildung 1 zeigt eine einfache Transformation, die aus einem Input (Shapefile) und zwei Outputs (PostgreSQL-Tabellen) besteht. Zwischen Input und Output

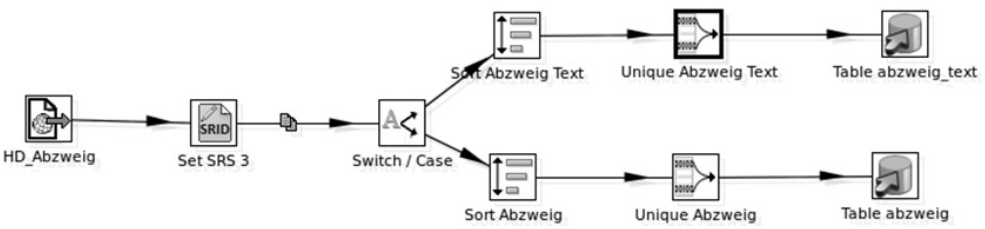

Abb. 1: Transformation

Strobl, J., Blaschke, T., Griesebner, G. & Zagel, B. (Hrsg.) (2014): Angewandte Geoinformatik 2014.

befinden sich weitere Steps, in denen die Daten verarbeitet werden. Verbindungen zwischen Steps werden Hop genannt. Die in den Transformationen definierten Schritte (Steps) können im Gegensatz zu den Jobs gleichzeitig ablaufen. Die Jobs sind den Transformationen übergeordnet. Sie kontrollieren den Ablauf von Transformationen und/oder anderen Jobs. Die Steps eines Jobs werden immer nacheinander abgearbeitet. Die Abbildung 2 zeigt einen einfachen Job. Zu Beginn eines Jobs steht immer der Start-Step. Danach folgen in diesem Job ein Step („Set varaiables 1 2“) sowie eine Transformation („get Subfolder“) und ein weiterer Job („Dateiimport Gas Step2“)

Abb. 2: Einfacher Job

2 Ausgangslage des Anwendungsfall

Bei einem Energiedienstleister in Thüringen, stand aufgrund eines Systemwechsels im Bereich des CAD auch ein Umzug der Geodaten von einer Oracle-Datenbank in eine PostgreSQL-Datenbank mit PostGIS an. Die Daten lagen in einem nicht dokumentierten Datenmodell vor und konnten nur mithilfe eines CAD als Shapedateien exportiert werden. Nach dem Export lagen 247 Shapefiles für die unterschiedlichen Objekte vor. Das Versorgungsgebiet des Energiedienstleisters ist allerdings so groß, dass der Export nur in 40 sich zum Teil überlagernden räumlichen Einheiten durchgeführt werden konnte. Die Daten dieser 40 Gebiete wurden als gezippte Dateien übergeben. Insgesamt wurden für das komplette Versorgungsgebiet ca. 11.000 Shapefiles erstellt. Die Inhalte der Shapefiles sind zudem nicht eindeutig, sodass ein Shapefile für Leitungen sowohl die Leitungen mit Ihren Attributen als auch Hilfslinien für die Beschriftungen mit den gleichen Attributen beinhaltet.

3 GeoKettle in action

Beim Import der Daten in die PostgreSQL-Datenbank mussten unter anderem folgende Arbeiten vorgenommen werden:

- Import der 11.000 Shapefiles in zwei verschiedene Datenbanken (Gas und Grundkarte),
- Verteilen von Objekten in einem Shapefile auf mehrere Tabellen,
- Zusammenfassen von Objekten aus mehreren Shapefiles mit unterschiedlichen Attributfeldern in einer Tabelle,
- Veränderung von Attributen,
- Entfernen von redundanten Daten,
- …

3.1 Vorgehensweise

Der Gesamtimport wurde in zwei Phasen (Erstimport und Nachbereitung) aufgeteilt. Für eine bessere Übersichtlichkeit und größere Kontrollmöglichkeiten des Importprozesses wurde der Gesamtimport in zwei thematisch voneinander getrennte Modelle aufgeteilt, die in ihrem generellen Ablauf keine Unterscheide aufweisen. Somit gibt es ein Importmodell für die Daten, die in die Datenbank „Gas“ fließen und ein Modell für die Daten der Datenbank „Grundkarte“. Zusätzlich wurden die räumlichen Exporte in Gruppen zu maximal sechs Einheiten zusammengefasst und in insgesamt neun Ordner abgelegt.

3.2 Erstimport

Zu Beginn der Modellierung des Erstimports wurden die Transformationen für die 247 Shapefiles erstellt, obwohl sie im Modell zuletzt ausgeführt werden. Die Abbildung 1 zeigt ein typisches Beispiel für eine Transformation des Erstimports. Die Schritte dieser Transformation umfassen die Zuweisung des Koordinatensystems, ggfs. die Verteilung des Shapefiles auf verschiedene Tabellen und die Beseitigung von Duplikaten.

Der Ablauf des Erstimports stellt sich wie folgt dar. Der erste Job des Modells ist in Abbildung 2 zu sehen. In dem ersten Step „Set variables“ wurde die Variable „root.path“ mit dem Pfad zum Arbeitsordner gesetzt, in welchem sich die Unterordner mit den zu importierenden Shapefiles befinden. In dieser Transformation „get Subfolders“, die in Abbildung 3

Abb. 3: Transformation „Get subfolders“

zu sehen ist, wurden die Unterordner des in der Variablen bestimmten Ordners ausgelesen und verschiedene Parameter mit dem Schritt „Copy rows to result“ an den folgenden Job „Datenimport Gas Step 2“ der in Abbildung 4 zu sehen ist übergeben. In diesem Job wurden aus den Parametern in der Transformation „getdirectories“ die Ordnernamen ausgelesen und nacheinander an die folgenden Jobs übergeben. In diesen Jobs wurde dann mithilfe des übergebenen Ordnernamens mit dem Step „Resultlist for Unzipped Files“ eine Liste der vorhandenen Shapefiles erstellt. In dieser Liste wurde überprüft, ob bestimmte Shapefiles vorhanden sind. Wenn die nachgefragte Datei vorhanden war, wurde die entsprechende Transformation für den Import durchgeführt, wenn nicht, wurde nach der nächsten Datei in der Liste geschaut. Dieses Vorgehen ist in Abbildung 5 exemplarisch für die Gruppe Abzweig dargestellt.

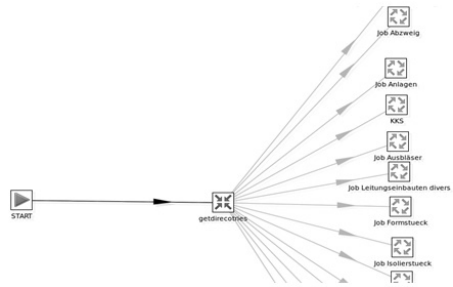

Abb. 4:
Job „Dateiimport Gas Step2“

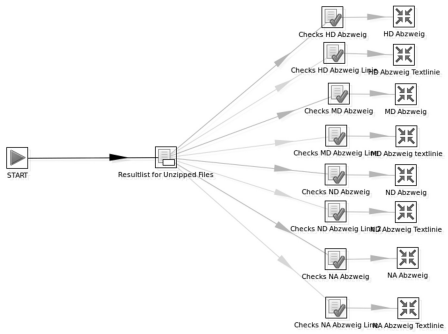

Abb. 5:
Job „Job Abzweig“

3.3 Nachbereitung

In der zweiten Phase ging es dann hauptsächlich darum, die Duplikate, die durch die räumlich Überlagerung der Importbereiche in die Datenbank geflossen sind, zu entfernen. Zudem wurden während dieser Transformationen auch Veränderungen an Attributen vorgenommen. Die Abbildung 6 zeigt eine typische Transformation dieser Phase. Hier wurden

Abb. 6: Transformation der Phase 2 für Abzweige

zum Beispiel Attribute verändert, Werte umgerechnet und Duplikate entfernt. Der Step „Replace in string“ wurde in diesem Fall dazu verwendet, um im Attributfeld „stil“ den string „Style-“ zu entfernen. Die beiden nächsten Schritte „Add constants factor rad2grad“ und „Calculator rad2grad“ wurden dazu verwendet, den Drehungswinkel von Symbolen und Schriften, der in allen Fällen in Rad vorlag, nach Grad umzurechnen. Dazu wird im ersten der beiden genannten Schritte der Umrechnungsfaktor als Konstante gesetzt. Im darauf folgenden Schritt (vgl. Abbildung 7) werden dann das Attributfeld (Field A) und die

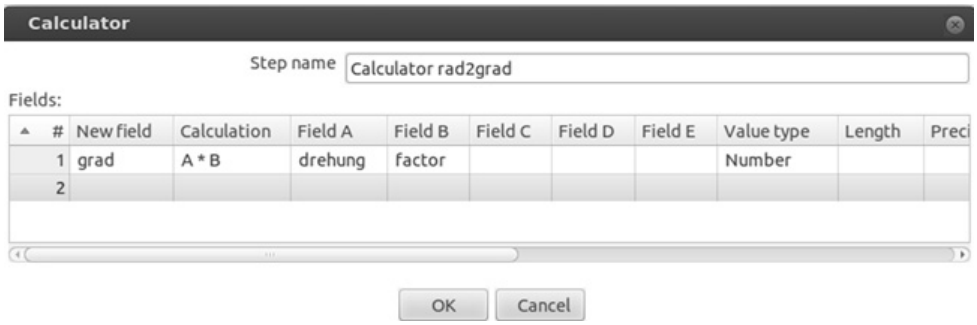

Abb. 7: Calculator

Konstante (Field B) ausgesucht und die Art der Berechnung angegeben (Calculation). Das Ergebnis wird in einem neuen Feld ausgegeben, das den Namen „grad“ erhält. Dieses neue Feld wird dann im Output, Step „Table abzweig unique“ auf das gewünschte Feld in der Datenbanktabelle gemappt. In dem vorletzten Schritt werden die zu importierenden Daten in eine Reihenfolge gebracht. Die Felder, nach denen sortiert werden soll, können dabei frei gewählt werden. In diesem Schritt gibt es zudem die Möglichkeit, Duplikate direkt zu entfernen. Die Sortierung ist dabei zwingend erforderlich, weil der aktuell zu bearbeitende Datensatz mit dem vorherigen Datensatz verglichen wird. In dem Fall, dass dieser Datensatz dem vorherigen entspricht, wird dieser nicht an den nächsten Step weiter geleitet.

Im Anschluss an die Nachbereitung wurden noch ein paar Schritte wie zum Beispiel das Anlegen der Primär- und Fremdschlüssel sowie die Erstellung eines räumlichen Index direkt auf der Datenbank durchgeführt. Diese Arbeiten hätte man aber auch ohne Weiteres noch in die Modellierung mit einbauen können.

4 Fazit

Ohne den Einsatz von GeoKettle wäre die Datenmigration kaum zu stemmen gewesen. Die ca. 11.000 Shapefiles wurden auf 2 Datenbanken mit insgesamt ca. 150 Tabellen verteilt. Die Modellierung des Importprozesses hat insgesamt etwa 2 Tage gedauert und der reine Importprozess keine 2 Stunden. Nach Überprüfung der importierten Daten war es sehr einfach, das Importmodell anzupassen, um Fehler zu beheben bzw. Ergänzungen vorzunehmen. Nach der Anpassung konnte dann entweder der komplette Import wiederholt oder nur die relevanten Daten neu importiert werden, indem die Hops zu nicht-relevanten Transformationen bzw. Jobs deaktiviert wurden.

GeoKettle wird in der WhereGroup in mehreren Projekten eingesetzt. So wird es zum Beispiel bei der Synchronisation von Datenbanken oder beim Mergen von Coderepositories eingesetzt. GeoKettle hat sich als gut dokumentierte und leicht zu erlernende Software heraus gestellt, die eine echte Alternative zur FME darstellt.