OpenSchoolMaps: Apache Hop mit GIS

OpenSchoolMaps.ch — Freie Lernmaterialien zu freien Geodaten und Karten

Ein Arbeitsblatt für Interessierte und Studierende

Einleitung

Apache Hop (Hop Orchestration Platform) ist eine Open-Source-Plattform zur Datenintegration. Sie

unterstützt sowohl die Daten-Orchestrierung (Workflows) als auch die Daten-Transformation (Pipelines). Es stehen verschiedene Erweiterungen (Plugins) zur Verfügung, darunter auch für die Transformation geografischer Daten - ähnlich wie Feature Manipulation Engine. Mithilfe von Hop können Nutzer Daten aus verschiedenen Quellen und Formaten verbinden, umwandeln und automatisieren, um sie anschliessend flexibel weiterzuverwenden oder zu analysieren. Die Software ermöglicht die Erstellung komplexer Datenintegrationsprozesse über eine grafische Benutzeroberfläche und deren automatisierte und überwachte Ausführung – ganz ohne Programmierkenntnisse. Die Plattform ist skalierbar und kann lokal (als Desktop- oder Serveranwendung) oder in der Cloud auf verschiedenen Betriebssystemen oder in Container-Umgebungen eingesetzt werden.

Als Einführung in die Software siehe das freie Arbeitsblatt auf OpenSchoolMaps zu "Daten sichten, bereinigen und integrieren mit Apache Hop".

Ziele

Dieses Arbeitsblatt zeigt, wie man das GIS-Plugin für Apache Hop in Pipelines verwenden kann, um verschiedene Geodaten-Formate einzulesen, anzureichern und wieder zu exportieren.

In diesem Arbeitsblatt lernen Sie:

- Das GIS-Plugin für Apache Hop zu installieren.
- Geodaten mit Apache Hop zu verarbeiten.
- Geodatenformate mit Apache Hop einzulesen und zu exportieren.
- Die wichtigsten GIS-Transforms in Apache Hop zu verstehen und anzuwenden.

Zeitplanung

Die Installation des GIS-Plugins dauert etwa eine Viertelstunde. Das reine Durchlesen dieses Arbeitsblatts dauert maximal eine halbe Stunde. Für das Lösen der Aufgaben ist zusätzlich etwa eine Stunde einzuplanen. Alle Angaben können je nach Einzelfall abweichen.

Voraussetzungen

Um das Arbeitsblatt durcharbeiten zu können, brauchen Sie folgende Dinge:

- Internetzugang zum Herunterladen von der benötigten Software und Daten.
- Java Version 17, verfügbar bei OpenJDK (Aufgrund von rechtlichen Unsicherheit raten öffentliche Stellen von der Nutzung von Oracle Java ab)
- Software: Apache Hop Version 2.12 (oder neuer), verfügbar für Windows, Mac und Linux, und zu installieren wie unten beschrieben.
- Software: GIS-Plugin für Apache Hop installiert, wie unten beschrieben.
- Daten: Datei "HOP-GIS-Daten.zip" von OpenSchoolMaps.

Um dieses Arbeitsblatt zu verstehen, sollten Sie Apache Hop bereits kennen. Eine Einführung bietet

das oben erwähnte, frei verfügbare Arbeitsblatt „Daten sichten, bereinigen und integrieren mit Apache Hop”.

Installation des Apache Hop GIS-Plugins

Im Folgenden wird Schritt für Schritt erklärt, wie das Apache Hop GIS-Plugin installiert werden kann. Der Vorgang kann entweder mit einem Script oder manuell durchgeführt werden.

Installation mit Script

Im Folgenden wird das Script gezeigt, welches verwendet wird, um das Plugin zu installieren. Es müssen vor dem Ausführen jedoch noch einige Anpassungen vorgenommen werden. Verwenden Sie dazu einen Texteditor.

- `HOP_GIS_PLUGIN_VERSION` mit der aktuellsten Versionsnummer, welche auf der GitHub-Seite eingesehen werde kann ersetzen.
- `HOP_HOME` mit dem Installationspfad von Apache Hop ersetzen. (Der Ordner, in welchem sich auch `hop-gui.bat` befindet.)

```
REM 1. Setze die Variablen
SET HOP_GIS_PLUGINS_VERSION="1.3.1"
SET HOP_HOME="C:\Program Files\Apache_Hop\2.12.0"

REM 2. Wechsle ins Download-Verzeichnis für den Download
cd C:\Users\%USERNAME%\Downloads

REM 3. Lade die Zip-Datei herunter
curl -L -o gis-plugin-assemblies-%HOP_GIS_PLUGINS_VERSION%.zip
https://github.com/atolcd/hop-gis-
plugins/releases/download/v%HOP_GIS_PLUGINS_VERSION%/gis-plugin-assemblies-
%HOP_GIS_PLUGINS_VERSION%.zip

REM 4. Kontrolle ob die zip-Datei vorhanden ist
dir gis-plugin-assemblies-%HOP_GIS_PLUGINS_VERSION%.zip

REM 4. Installiere die Dateien ins Plugin-Verzeichnis
tar -xf gis-plugin-assemblies-%HOP_GIS_PLUGINS_VERSION%.zip -C %HOP_HOME%\plugins

REM 5. Kontrolle ob die Plugin-Dateien vorhanden sind
dir %HOP_HOME%\plugins\gis
```

Zunächst muss eine CMD-Shell geöffnet werden. Hierbei empfiehlt es sich, auf Windows **nicht** Powershell zu verwenden.

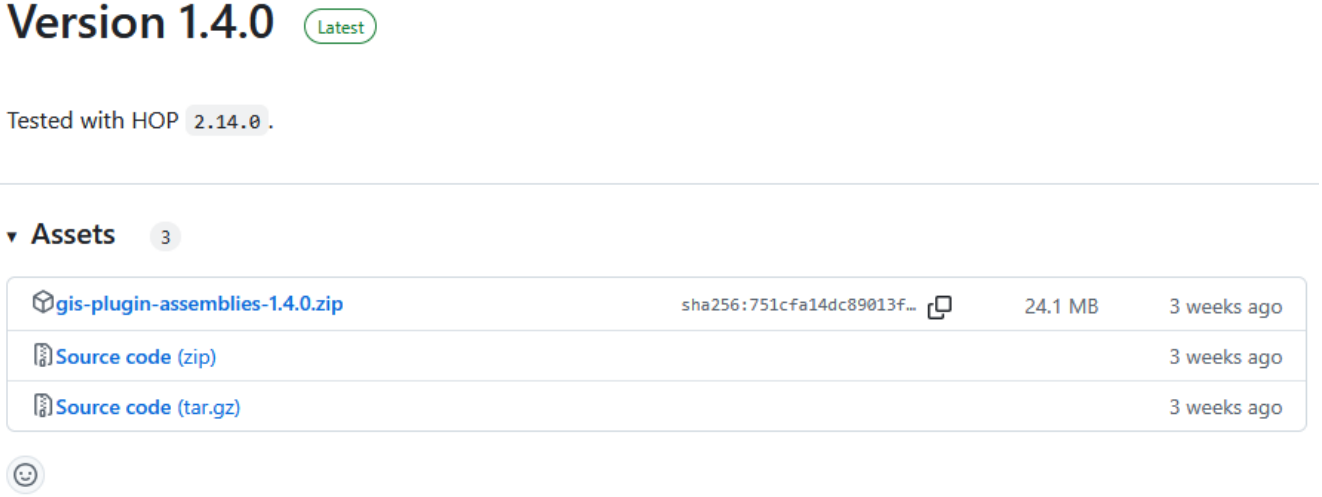

Abbildung 1. Die GitHub-Release Seite mit der zum Zeitpunkt der Erstellung neusten Version.

Manuelle Installation

Alternativ kann das GIS-Plugin auch manuell installiert werden. Auf der GitHub-Seite kann die neuste Version heruntergeladen werden.

Struktur des Arbeitsblatts

- Einführung: Überblick über Apache Hop und das GIS-Plugin, inklusive Ziele, Voraussetzungen und Installationsanleitung.
- Grundlagen des GIS-Plugins: Beschreibung der wichtigsten GIS-spezifischen Transforms und ihrer Anwendung in Apache Hop.
- Übungen: Schritt-für-Schritt-Anleitungen zur Verarbeitung, Anreicherung und Exportieren (Burgen mit QRank anreichern) sowie das räumliche Einordnen (Sushi-Restaurants filtern) von Geodaten mit Apache Hop.
- Abschluss: Zusammenfassung der wichtigsten Erkenntnisse und Reflexion über die erlernten Fähigkeiten.
- Was gelernt wurde: Übersicht der wichtigsten Lektionen aus den Übungen und ihrer praktischen Relevanz.

GIS-Plugin-Grundlagen

Das GIS-Plugin, das aus sieben Teil-Plugins besteht, erweitert Apache Hop um spezialisierte Funktionen zur Verarbeitung räumlicher Daten. Damit lassen sich Geodaten in verschiedenen Formaten nicht nur einlesen und exportieren, sondern auch analysieren, transformieren und in Beziehung zueinander setzen. Das Plugin ist besonders für Aufgaben im Bereich Data Engineering und Data Science mit geografischem Bezug ein mächtiges Werkzeug, da es typische GIS-Funktionalitäten wie Koordinatentransformationen, Geometrieberechnungen oder räumliche Beziehungen direkt in Pipelines integrierbar macht. In diesem Abschnitt werden die wichtigsten Transforms des GIS-Plugins mit praktischen Hinweisen vorgestellt.

Auf GitHub gibt es ebenfalls eine Dokumentation, in welcher auf die einzelnen Transforms eingegangen wird.

GIS Transforms

Das GIS-Plugin fügt Apache Hop einige neue Transforms und einen Datentyp hinzu, die den Umgang mit Geodaten ermöglichen und vereinfachen. Im Folgenden werden diese Transforms erläutert und ihre Stolpersteine aufgezeigt.

Dateipfade, die in den Transforms des GIS-Plugins eingetragen werden, dürfen keine Leerzeichen enthalten. Andernfalls findet Apache Hop die Datei nicht.

Transform GIS File input

Durch den `GIS File input`-Transform lassen sich verschiedene Geodaten-Formate einlesen lassen. Darunter: `GeoPackage`, `GeoJSON`, `MapInfo Interchange Format (MIF)`, `SpatialLite SQLite`, `ESRI Shapefile`, `DXF (AudoCAD)`, und `GPS eXchange Format (GPX)`.

Formate wie GeoJSON erlauben es Objekten, unterschiedliche Felder zu enthalten ("schemaless"). Hop benötigt jedoch eine Schemadefinition (d.h. Feldbeschreibungen), die für alle Objekte in einem Eingabestrom gilt. Daher muss `GIS File input` die Schemadefinition aus den Daten extrahieren. Dazu werden die Eigenschaften der GeoJSON-Objekte verwendet.

Transform GIS File output

Durch den `GIS File output`-Transform lassen sich Daten in verschiedenen Geodatenformaten abspeichern bzw. exportieren; darunter: `GeoPackage`, `GeoJSON`, `ESRI Shapefile`, `DXF (AutoCAD)`, `GPS eXchange Format (GPX)`, `Keyhole Markup Language (KML)` und `Scalable Vector Graphics (SVG)`.

Beim Shapefile Export wird die .prj-Datei nicht geschrieben. Damit erkennen nachfolgende Werkzeuge das zugehörige Koordinatenreferenzsystem (CRS) der Daten nicht.

Um Geopackages zu schreiben, muss ein `EPSG code` angegeben werden, auch wenn dieser nicht als *Required* gelistet ist. Und: Eine GeoPackage-Datei (evtl. auch andere Formate?) kann nicht überschrieben werden, da der Java-Prozess die Datei nicht freigibt.

Transform Coordinate transformation

Durch den `Coordinate transform`-Transform lassen sich Geometrien von einem Koordinatenreferenzsystem (CRS) in ein anderes überführen, z.B. von `EPSG:4326` (WGS84) nach `EPSG:2056` (LV95 Schweiz). Der Schritt erwartet ein gültiges Geometriefeld (z.B. aus einem vorgelagerten GIS-Schritt) sowie die Angabe von Quell- und Ziel-CRS. Die Transformation erfolgt auf Basis anerkannter EPSG-Codes und ermöglicht eine nahtlose Weiterverarbeitung in verschiedenen

räumlichen Bezugsrahmen.

Transform Geometry information

Durch den `Geometry information`-Transform lassen sich aus einem Geometriefeld zusätzliche räumliche Eigenschaften extrahieren. Dazu zählen unter anderem: `Fläche`, `Länge`, `Geometrietyp`, `SRID` (EPSG-Code) sowie Anzahl der Punkte. Der Schritt eignet sich besonders zur Analyse, Filterung oder Qualitätssicherung geografischer Daten innerhalb eines Hop-Workflows.

Transform Geoprocessing

Mit dem `Geoprocessing`-Transform können verschiedene Operationen an Geodaten vorgenommen werden. Darunter fallen unter anderem: `Get coordinates`, `Simplify`, `Union`, `Longest line`, `Largest polygon`, `Intersection`, `Invers` und viele weitere. Eine Liste aller Operationen kann in der Dokumentation auf GitHub gefunden werden. In den Parametern muss jeweils der Operator und die betroffenen Geometrien ausgewählt werden.

Transform Geospatial Group by

Der `Geospatial Group by`-Transform kann genutzt werden, um Einträge anhand räumlicher Attribute zu gruppieren. Der Transform findet dabei z.B. beim Aussortieren von Duplikaten Anwendung.

Transform Spatial relationship

Durch den `Spatial relationship`-Transform kann überprüft werden, in welcher Beziehung zwei Objekte stehen. Ob z.B. ein Objekt in einem bestimmten Polygon steht (`within`) oder ob ein Objekt ein anderes Berührt (`touches`).

Die folgenden Übungen enthalten nebst den Aufgaben auch Hinweise auf Stolpersteine. Im Arbeitsblatt "Daten sichten, bereinigen und integrieren mit Apache Hop" auf OpenSchoolMaps gibt es auch "Best Practices".

Übungen

Übung 1: Burgen mit QRank anreichern

Im Kartenprojekt OpenStreetMap(OSM) sind Burgen und Ruinen umfassend erfasst. In der Regel besitzen sie einen Verweis auf Wikidata (sogenanntes Q-Element). Wikidata ist eine frei zugängliche, gemeinschaftlich gepflegte und mehrsprachige Datenbank. Sie dient als zentraler Knotenpunkt für sämtliche Wikipedia-Sprachversionen und andere Wikimedia-Projekte.

Was vielen praktischen Anwendungen jedoch fehlt, ist eine sinnvolle Priorisierung der Objekte nach Relevanz. Hier kommt der freie Datensatz QRank ins Spiel: Für jedes Q-Element in Wikidata stellt er einen vorab berechneten Ranking-Wert bereit. Dieser Wert basiert auf den Seitenaufrufen der jeweiligen Wikipedia-Artikel. Je häufiger ein Artikel aufgerufen wird, desto höher wird die Bedeutung des zugehörigen Objekts eingeschätzt.

Daten

Für diese Aufgabe werden die aktuellsten Daten von QRank und das `Castles_CH.geojson` aus der ZIP-Datei benötigt. Das qrank.csv kann hier heruntergeladen werden.

Schritt 1.1: Daten einlesen

In einem neuen Projekt müssen zunächst die Daten eingelesen werden, um sie weiterzuverarbeiten. Verwenden Sie hierfür den `GIS Import`-Transform. In diesem Transform muss lediglich der Dateityp und der Pfad zur Datei auf das entsprechende `Castles_CH.geojson` angepasst werden.

Das `qrank.csv` kann hingegen mit einem bereits vorinstallierten `CSV file input`-Transform eingelesen werden.

Schritt 1.2 Beide Datenströme sortieren

Um beide Tabellen zusammenzuführen, müssen diese zunächst sortiert werden. Fügen Sie hierfür einen `Sort rows`-Transform hinter beiden `File Inputs` ein. Konfigurieren Sie diese so, dass beide Tabellen aufsteigend entsprechend der `Wikidata`-Spalte sortiert werden.

Schritt 1.3 Burgen anreichern

Nun können beide Datenströme mit einem `Merge join`-Transform verbunden werden. Da im finalen Datensatz die Tabelle mit den Burgen (also `Castles_CH.geojson`) mit dem Rank angereichert werden soll, verwenden Sie einen `LEFT OUTER`-Join, wobei die Tabelle mit den Burgen die erste Eingabe und die Tabelle mit dem QRank die zweite Eingabe darstellt.

Schritt 1.4 Daten speichern

Im nächsten Schritt können die Daten mit einem `GIS File output`-Transform als GeoJSON abgespeichert werden. Passen Sie hierfür den `Type` und den `Filename` an.

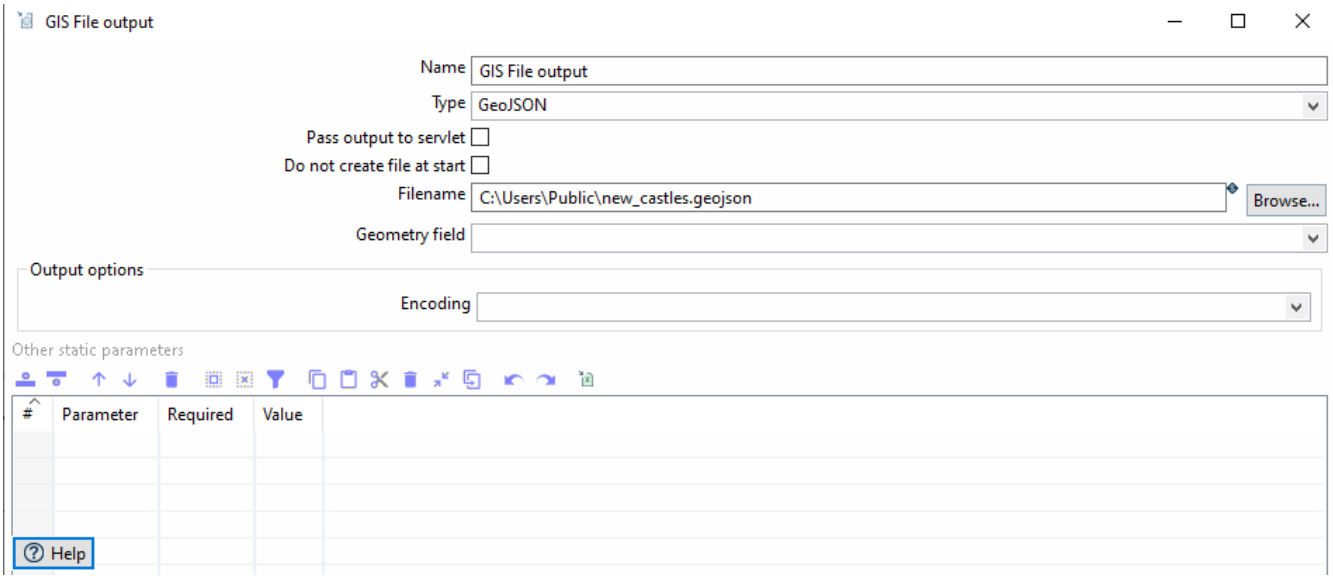

Abbildung 2. Die angepasste Konfiguration, um die Daten als GeoJSON zu exportieren.

Schritt 1.5 Daten überprüfen

Zur Überprüfung des Exports öffnen Sie die Datei z.B. mit geojson.io oder QGIS.

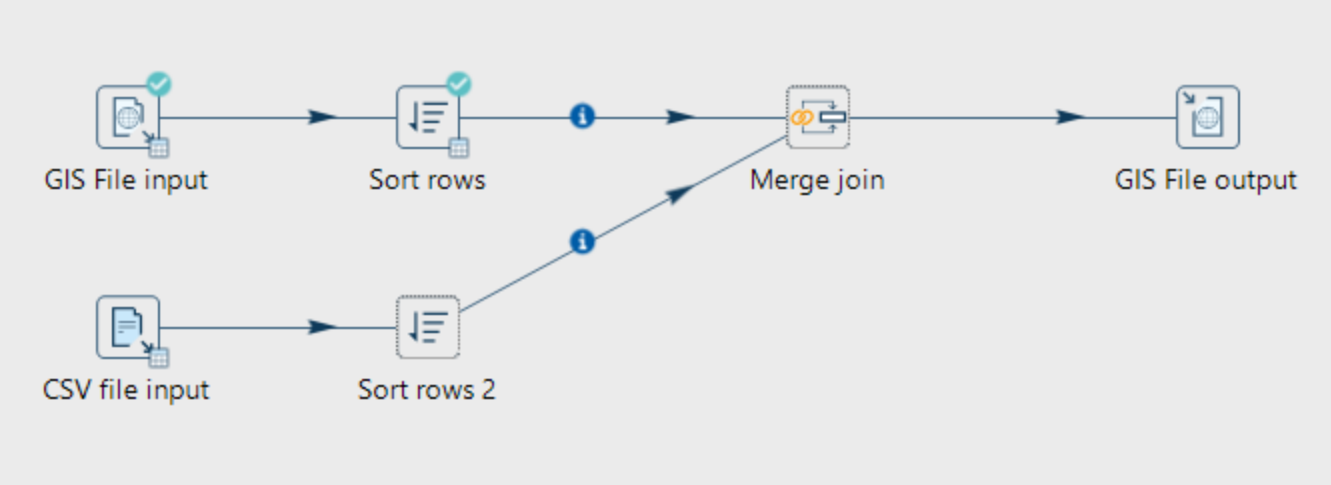

Abbildung 3. Die finale Pipeline der ersten Übung.

Übung 2: Sushi-Restaurants filtern

Ziel dieser Aufgabe ist es, aus einem Datensatz, aller auf OSM eingetragener Sushi-Restaurants, nur Restaurants mit einem Outdoor-Bereich herauszufiltern sowie deren Adresse zusammenzuführen. Zudem werden die Daten auf den Raum der Schweiz eingegrenzt.

Daten

Gegeben ist ein GeoJSON `sushirestaurants.geojson` mit besagten Restaurants. Zudem wird das aktuellste GeoJSON des Schweizer Landesgebiets benötigt.

Schritt 2.1 Daten einlesen

Wie zuvor auch, kann das GeoJSON mithilfe von einem `GIS File input`-Transforms eingelesen werden.

Schritt 2.2 Daten anpassen

Sind die Daten eingelesen, können sie mithilfe von einem `Select Values`-Transform angepasst werden. Die Spalten `@id`, `addr:street`, `addr:housenumber`, sollen dabei umbenannt werden, so dass sie keine Sonderzeichen '@' und ':' enthalten.

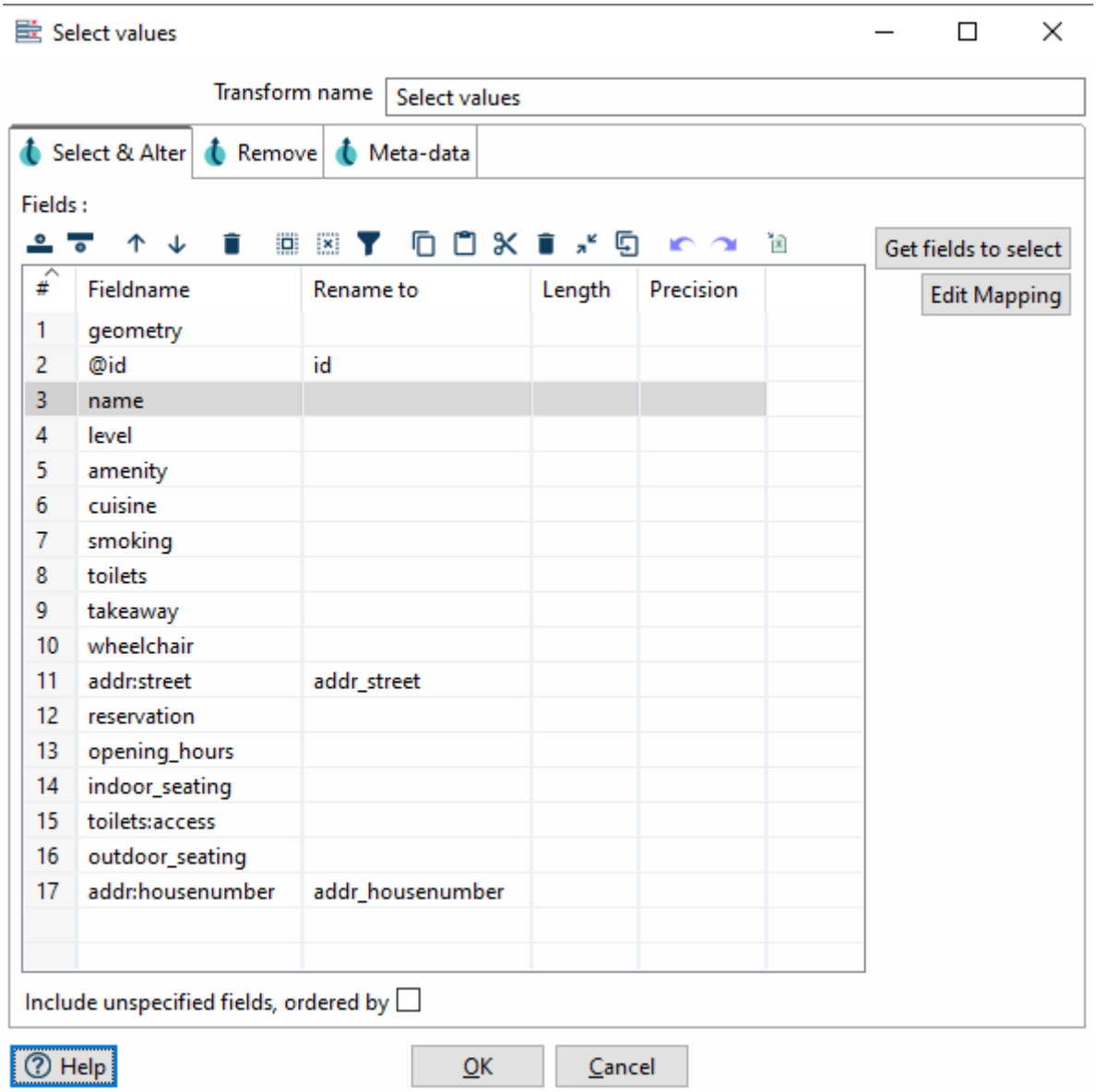

Abbildung 4. Die Konfiguration des Select Values-Transforms

Schritt 2.3 Nach Outdoor Restaurants filtern

Mithilfe von dem `Filter rows`-Transform kann anschliessend nach den Outdoor-Restaurants gefiltert werden. Hierzu muss in der `condition` nach `outdoor_seating` geprüft werden.

Schritt 2.4 Adresse zusammenführen

Nun kann die Adresse, die aktuell noch aufgeteilt auf Strasse und Hausnummer ist, zusammengeführt werden. Dafür kann der `Concat`-Transform verwendet werden. Im `Concat` -Transform muss ein entsprechendes Target-Filed `addr_full` und ein entsprechender Seperator `|` konfiguriert werden. Anschliessend können die Felder, welche zusammengeführt werden sollen im Tab `Fields`, ausgewählt werden.

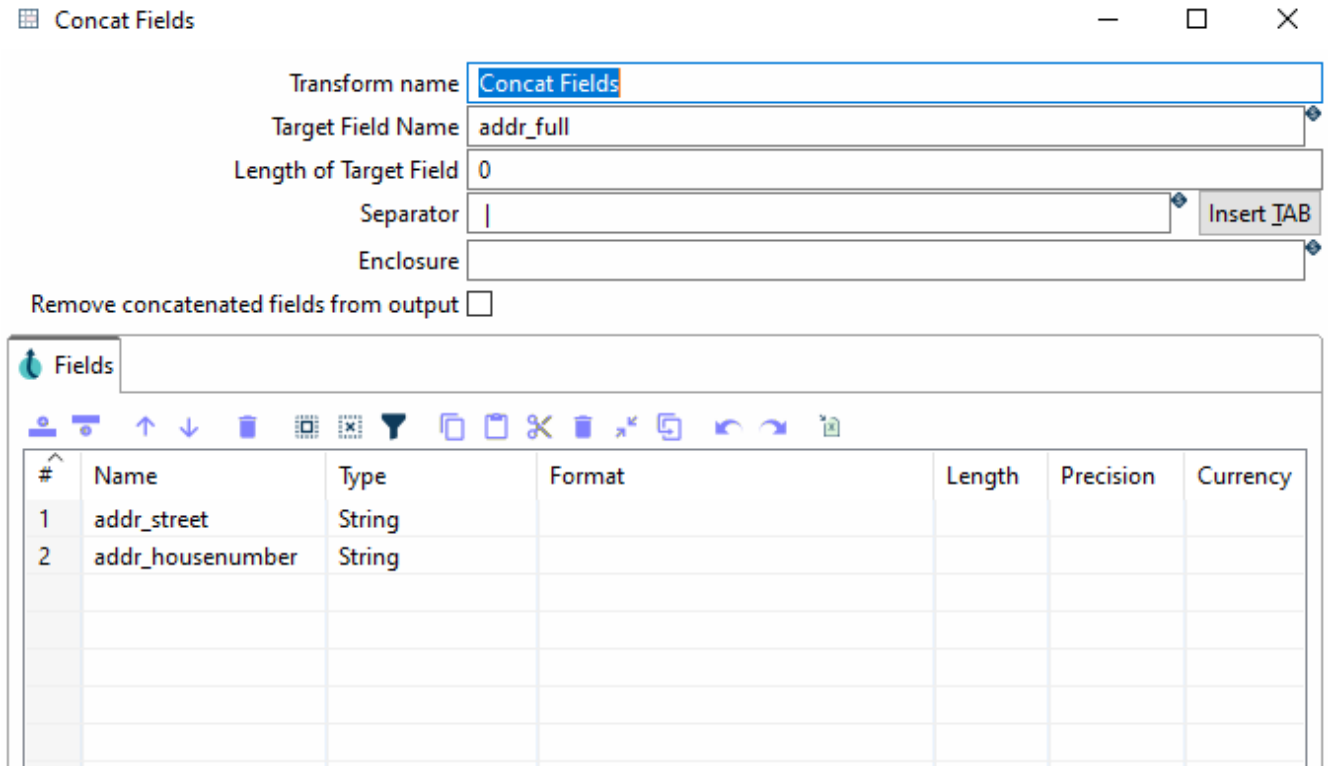

Abbildung 5. Die Konfiguration des Concat-Transforms

Schritt 2.5 Daten Räumlich eingrenzen

Sobald die Adresse zusammengeführt wurde, kann ein neuer Ast angelegt werden, in welchem zunächst die Daten der Schweizer Landesgrenzen aufbereitet und anschliessend in die Daten eingebunden wird.

Zunächst kann mit einem `Gis File input`-Transform das `swissBOUNDARIES3D_1_3_TLM_LANDESGEBIET.geojson` eingelesen werden.

Anschliessend können die Daten mit einem `Filter Rows`-Transform auf die Daten der Schweizer Grenze begrenzt werden. Hierzu kann das Feld `NAME` gegen den Value `Schweiz`, getestet werden.

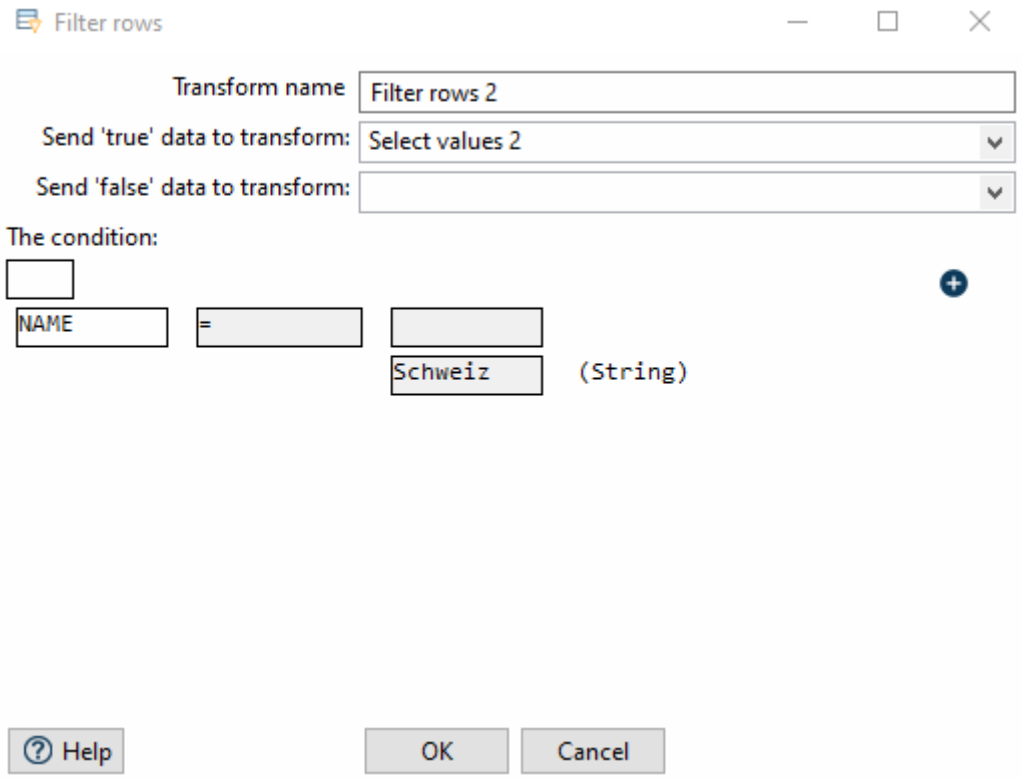

Abbildung 6. Filter mit den Parametern für die Schweizer Grenze

Nun können alle Spalten, welche nicht benötigt werden, mit einem `Select Values`-Transform entfernt werden. Benötigt wird nur das Feld mit der Geometrie (`geometry`).

Somit wurden die Daten auf die Geometrie der Schweizer Grenze eingegrenzt, welche indessen genutzt werden können, um zu prüfen, ob sich ein Restaurant in der Schweiz befindet. Diese Daten können nun mit einem `Join Rows (cartesian product)`-Transform mit den Daten der Sushi-Restaurants angereichert werden. Wichtig ist, dass zunächst der Ast mit den Daten der Sushi-Restaurants angeschlossen wird und dann erst die Daten mit der Schweizer Grenze.

Die Daten müssen auf diese Weise zusammengeführt werden, da der `Spatial relationshop and proximity`-Transform nur einen Hop entgegennehmen kann.

Sobald die Daten zusammengeführt wurden, können diese in den `Spatial relationshop and proximity`-Transform geführt werden. Mit diesem wird getestet, ob sich ein Restaurant in der Schweiz befindet. Dafür muss der Transform wie folgt eingestellt werden. `Spatial relationship` sollte auf `within` umgestellt werden. `Geometry A` sollte das Feld `geometry` und `Geometry B` sollte auf `geometry_1` gestellt werden.

Nun können die Felder, welche im Output nicht mehr benötigt werden, mit einem `Select Value` entfernt werden.

Schritt 2.6 GeoPackage exportieren

Abschliessend können die Daten wie zuvor mit einem `GIS File export`-Transform exportiert werden. Diesmal soll jedoch ein GeoPackage exportiert werden. Hierfür sind einige Dinge zu beachten.

Um ein GeoPackage richtig exportieren zu können, wird eine Id-Spalte benötigt. Um diese zu erzeugen, kann ein `Add-Sequence`-Transform verwendet werden, welcher wie folgt konfiguriert

werden soll.

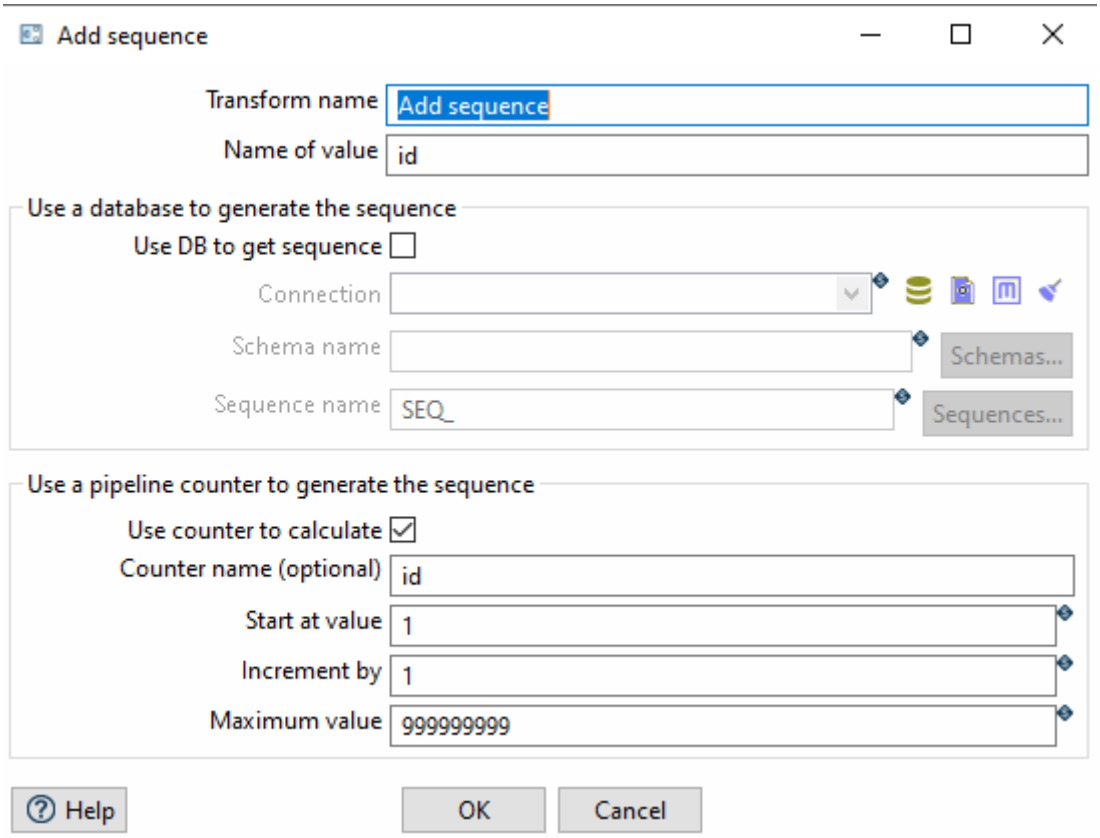

Abbildung 7. Konfiguration des Add Sequence-Transforms.

Anschliessend kann der `GIS File output`-Transform angeschlossen werden. Von diesem Transform müssen im Folgenden noch einige Anpassungen gemacht werden. Der `Type` muss auf GeoPackage geändert werden. Und es muss ein `Table name` und ein `EPSG code` angegeben werden.

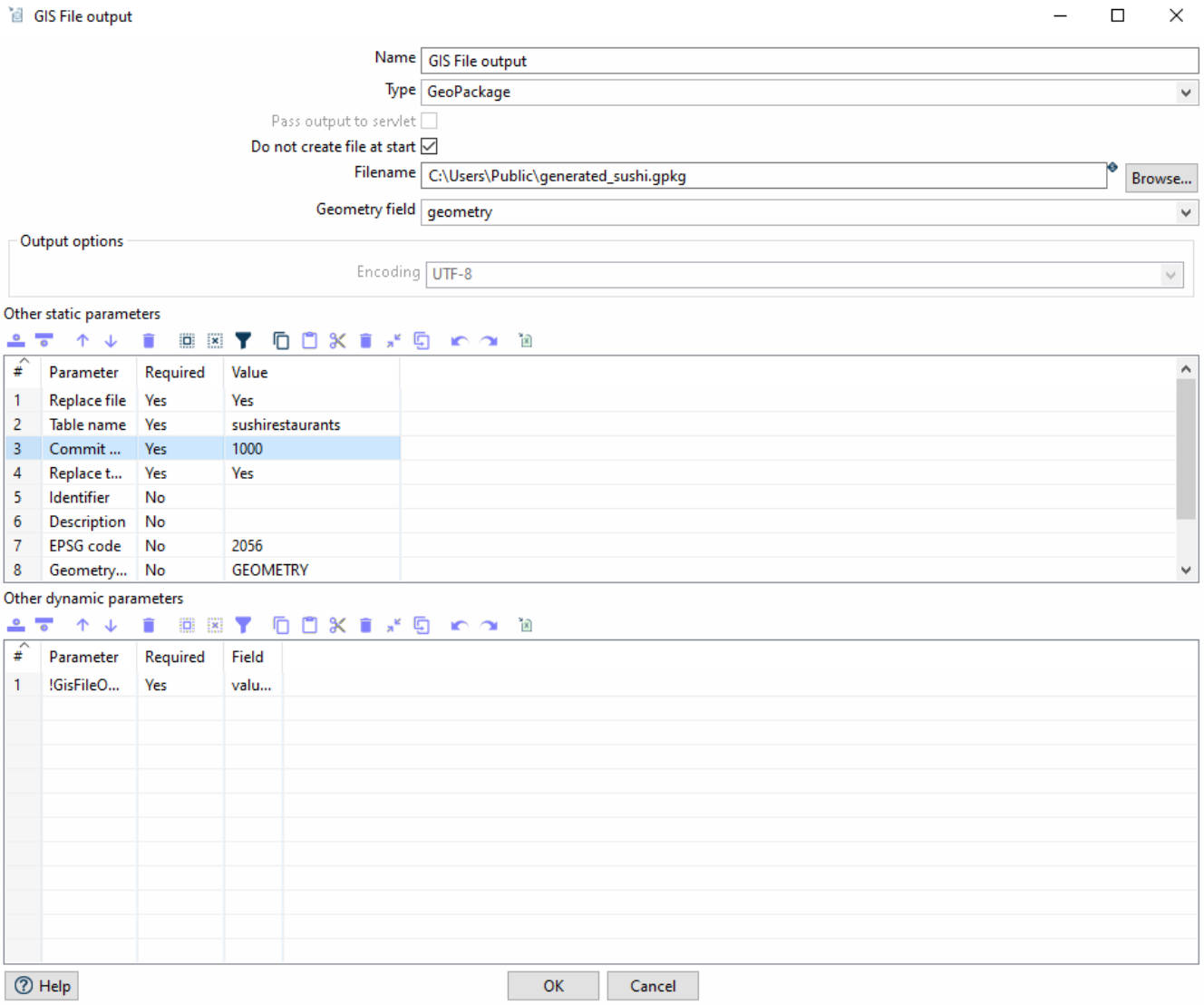

Abbildung 8. Die Konfiguration des GIS File output-Transforms.

Zudem kann der `Geometry type` auf `POINT` geändert werden, muss jedoch nicht.

Es kann dazu kommen, dass der `GIS File output`-Transform einen Error auslöst, wenn bereits eine Datei existiert. Darum ist es empfohlen, bereits existierende Dateien zu löschen, bevor man die Pipeline ausführt.

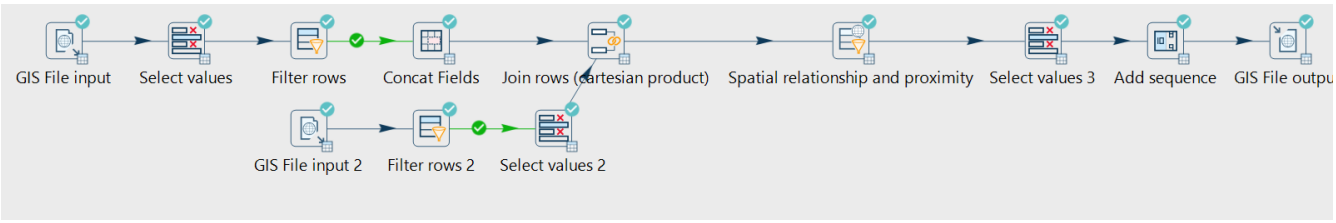

Abbildung 9. Die fertige Pipeline im ausgeführten Zustand.

Abschluss

Anhand der besprochenen Punkte und Übungen wurde deutlich, dass sich Apache Hop auch zur Integration und Verarbeitung von Geodaten eignet. Apache Hop basiert auf einer ausgereiften, professionellen Architektur. Viele Tools (z. B. Workflows) und Funktionen (z. B.

Konfigurationen/Metadaten) wurden noch nicht vorgestellt.

Dabei zeigte sich jedoch auch, dass das GIS-Plugin den Funktionsumfang von FME noch nicht erreicht. Apache Hop ist jedoch aufgrund seiner Open-Source-Natur relativ einfach und schnell erweiterbar. Tragen Sie zum Projekt bei, indem Sie Verbesserungsvorschläge oder Wünsche einbringen oder bei der Dokumentation mithelfen.

Was gelernt wurde

- Installation des GIS-Plugins für Apache Hop.
- Erstellung eines neuen Projekts in Apache Hop.
- Einlesen und Exportieren geografischer Daten mithilfe von GIS-spezifischen Transforms.
- Transformation und Anreicherung von Geodaten durch Kombinieren unterschiedlicher Datenquellen (QRank).
- Filtern von Daten basierend auf ihrer räumlichen Einordnung.
- Bereinigung und Filterung von Geodaten – z.B. durch Umbenennen von Attributen, Entfernen von Duplikaten oder bedingtes Filtern.
- Durchführung von Koordinatentransformationen.
- Erstellung neuer Attributwerte – etwa durch Adress-Zusammenführung oder sequentielle IDs.

Noch Fragen? Sehen Sie auch "Kontakt" auf OpenSchoolMaps!