Geophysical Research Abstracts
Vol. 18, EGU2016-13938-1, 2016
EGU General Assembly 2016

Investigation and Evaluation of the open source ETL tools GeoKettle and Talend Open Studio in terms of their ability to process spatial data

Kristin Kuhnert (1) and Jörn Quedenau (2)

(1) Institute for Advance Sustainability Studies, Potsdam, Germany, (2) Institute for Advance Sustainability Studies, Potsdam, Germany

Integration and harmonization of large spatial data sets is not only since the introduction of the spatial data infrastructure INSPIRE a big issue. The process of extracting and combining spatial data from heterogeneous source formats, transforming that data to obtain the required quality for particular purposes and loading it into a data store, are common tasks.

The procedure of Extraction, Transformation and Loading of data is called ETL process. Geographic Information Systems (GIS) can take over many of these tasks but often they are not suitable for processing large datasets. ETL tools can make the implementation and execution of ETL processes convenient and efficient. One reason for choosing ETL tools for data integration is that they ease maintenance because of a clear (graphical) presentation of the transformation steps. Developers and administrators are provided with tools for identification of errors, analyzing processing performance and managing the execution of ETL processes. Another benefit of ETL tools is that for most tasks no or only little scripting skills are required so that also researchers without programming background can easily work with it.

Investigations on ETL tools for business approaches are available for a long time. However, little work has been published on the capabilities of those tools to handle spatial data. In this work, we review and compare the open source ETL tools GeoKettle and Talend Open Studio in terms of processing spatial data sets of different formats.

For evaluation, ETL processes are performed with both software packages based on air quality data measured during the BÄRLIN2014 Campaign initiated by the Institute for Advanced Sustainability Studies (IASS). The aim of the BÄRLIN2014 Campaign is to better understand the sources and distribution of particulate matter in Berlin. The air quality data are available in heterogeneous formats because they were measured with different instruments. For further data analysis, the instrument data has been complemented by other georeferenced data provided by the local environmental authorities. This includes both vector and raster data on e.g. land use categories or building heights, extracted from flat files and OGC-compliant web services.

The requirements on the ETL tools are now for instance the extraction of different input datasets like Web Feature Services or vector datasets and the loading of those into databases. The tools also have to manage transformations on spatial datasets like to work with spatial functions (e.g. intersection, union) or change spatial reference systems.

Preliminary results suggest that many complex transformation tasks could be accomplished with the existing set of components from both software tools, while there are still many gaps in the range of available features. Both ETL tools differ in functionality and in the way of implementation of various steps. For some tasks no predefined components are available at all, which could partly be compensated by the use of the respective API (freely configurable components in Java or JavaScript).