Geophysical Research Abstracts
Vol. 18, EGU2016-11780, 2016
EGU General Assembly 2016

Scalable Earth-observation Analytics for Geoscientists: Spacetime Extensions to the Array Database SciDB

Marius Appel (1), Florian Lahn (1), Edzer Pebesma (1), Wouter Buytaert (2), and Simon Moulds (2)

(1) Institute for Geoinformatics, University of Muenster, Muenster, Germany, (2) Department of Civil and Environmental Engineering, Imperial College London, London, United Kingdom

Today's amount of freely available data requires scientists to spend large parts of their work on data management. This is especially true in environmental sciences when working with large remote sensing datasets, such as obtained from earth-observation satellites like the Sentinel fleet. Many frameworks like SpatialHadoop or Apache Spark address the scalability but target programmers rather than data analysts, and are not dedicated to imagery or array data. In this work, we use the open-source data management and analytics system SciDB to bring large earth-observation datasets closer to analysts. Its underlying data representation as multidimensional arrays fits naturally to earth-observation datasets, distributes storage and computational load over multiple instances by multidimensional chunking, and also enables efficient time-series based analyses, which is usually difficult using file- or tile-based approaches. Existing interfaces to R and Python furthermore allow for scalable analytics with relatively little learning effort. However, interfacing SciDB and file-based earth-observation datasets that come as tiled temporal snapshots requires a lot of manual bookkeeping during ingestion, and SciDB natively only supports loading data from CSV-like and custom binary formatted files, which currently limits its practical use in earth-observation analytics. To make it easier to work with large multi-temporal datasets in SciDB, we developed software tools that enrich SciDB with earth observation metadata and allow working with commonly used file formats: (i) the SciDB extension library *scidb4geo* simplifies working with spatiotemporal arrays by adding relevant metadata to the database and (ii) the Geospatial Data Abstraction Library (GDAL) driver implementation *scidb4gdal* allows to ingest and export remote sensing imagery from and to a large number of file formats. Using added metadata on temporal resolution and coverage, the GDAL driver supports time-based ingestion of imagery to existing multi-temporal SciDB arrays. While our SciDB plugin works directly in the database, the GDAL driver has been specifically developed using a minimum amount of external dependencies (i.e. CURL). Source code for both tools is available from github [1]. We present these tools in a case-study that demonstrates the ingestion of multi-temporal tiled earth-observation data to SciDB, followed by a time-series analysis using R and SciDBR. Through the exclusive use of open-source software, our approach supports reproducibility in scalable large-scale earth-observation analytics. In the future, these tools can be used in an automated way to let scientists only work on ready-to-use SciDB arrays to significantly reduce the data management workload for domain scientists.

[1] https://github.com/mappl/scidb4geo and https://github.com/mappl/scidb4gdal