Master Thesis Proposal

Streaming Web-Services for Calculating Live Hydrological Derivatives

Christian Autermann

September 27, 2013

Recent research has highlighted the relevance of lakes to global process such as the carbon cycle [1]. Ecological studies on lakes have historically taken advantage of the "closed system" bounds to delineate a simplified ecosystem, but analyses that are formulated to answer societally relevant questions often must scale this single system science approach to hundreds, thousands, or millions of lakes [2]. Therefore systems must be developed that can aggregate, analyze, and ultimately interpret hydrological data at large scales. Additionally, these analytical systems must be able to easily couple lake features with supporting data that define, for example, catchment properties, local climate, and anthropomorphic stressors. These data products are readily available as national coverages that can either be sampled and turned into model parameters, or turned into model drivers if they are time series products.

This work shall evaluate the existing tools (e.g. Lake Analyzer[1], see [3]), data models and the modeling frameworks used by USGS CIDA[2]. Modeling runs are based on online data brokers (such as the USGS's Geo Data Portal - GDP[3]) build upon Open Geospatial Consortium standards such as CSW[4], WPS[5] WMS[6] and WCS[7], but still rely on local algorithms, which comprise functionality for statistical quality assurance and quality control as well as the calculation of various metrics related to the physical state of the lakes (often linked with ecosystem function or disturbance). Building standardized and flexible infrastructure for analyzing foundational

[1] https://github.com/GLEON/Lake-Analyzer
[2] http://cida.usgs.gov/
[3] http://cida.usgs.gov/gdp/
[4] http://www.opengeospatial.org/standards/cat
[5] http://www.opengeospatial.org/standards/wps
[6] http://www.opengeospatial.org/standards/wms
[7] http://www.opengeospatial.org/standards/wcs

data used by domain scientists is an important challenge given legacy and heterogeneous architectures. Therefore building on the existing infrastructure and corresponding demands of the use case shall be considered.

One approach for a scalable system is to move the modeling to a web-based processing framework, which should rely on public and interoperable standards in the given use case. Web processing allows to chain data brokers with translators, models, and eventually post-hoc analysis of model runs. This chain provides specific information products to the user. Considering the amount of data (and future process scaling needs), such an analysis must be conducted in a streaming manner, i.e. the processing should start before the last chunk of data comes in, and the output should also be available in parts before the processing has completely finished to reduce the lag for domain users of the system. Existing approaches to this problem shall be critically evaluated.

This thesis work comprises the evaluation, design and prototypical implementation of a lake analysis chain for live sensor data. This includes the evaluation of existing datamodels (mainly CSV/TSV) and a standardized way to convert existing domain specific applications written in MatLab[8] into streaming web services (possibly WPS algorithms) in favor of the currently used non standardized web frontend[9].

Research Questions

- How can large scale hydrological data be processed in a service-based processing chain?
- Do available web-processing interface definitions support a live data streaming scenario, what is missing?
- Can real-time data be integrated into the processing chain for a constant (streamed) analysis?
- How does the developed architecture perform in practical test with 1000s and 10000s of lake features?
- How can continued statistical quality assurance and quality control in the application area of lake ecology be modeled in a web service chain?
- Do existing standards (data models and service interfaces for data warehousing, processing and visualization) support a streaming analysis chain? What is missing?
- How can a analysis language commonly used by domain experts (in this case MatLab) be easily deployed in a web based processing chain?
- How can spatial dependencies of features in streams be considered?

[8] http://www.mathworks.de/products/matlab/
[9] http://lakeanalyzer.gleon.org/

References

[1] J. J. Cole, Y. T. Praire, N. F. Caraco, W. H. Mcdowell, L. J. Tranvik, R. G. Striegl, C. M. Duarte, P. Kortelainen, J. A. Downing, J. J. Middelburg, J. Melack. 2007. *Plumbing the global carbon cycle: Integrating inland waters into the terrestrial carbon budget.* Ecosystems 10: 171 – 184.

[2] J. A. Downing, Y. T. Praire, J. J. Cole, C. M. Duarte, L. J. Tranvik, R. G. Striegl, W. H. Mcdowell, P. Kortelainen, N. F. Caraco, J. M. Melack, J. J. Middelburg. 2006. *The global abundance and size distribution of lakes, pounds, and impoundments.* Limnology and Oceanography 51: 2388 – 2397.

[3] J. S. Read, D. P. Hamilton, I. D. Jones, K. Muraoka, L. A. Winslow, R. Kroiss, C. H. Wu, E. Gaiser. 2011. *Derivation of lake mixing and stratification indices from high-resolution lake buoy data.* Environmental Modelling and Software 26: 1325 – 1336.

[4] J. S. Read, L. A. Winslow, G. A. Hansen, J. Van Den Hoek, C. D. Markfort, N. Booth. 2013. *Upscaling aquatic ecology: Linking continental data products to distributed lake models.* Poster, EarthCube Inland Waters meeting.