A Scalable Geoprocessing Workflow for Big Geo-Data Analysis and Optimized Geospatial Feature Conflation based on Hadoop

Song Gao, Linna Li, Michael F. Goodchild

Department of Geography, University of California, Santa Barbara, CA, USA

Keywords: Scalable geoprocessing workflow; Big Geo-data analysis; MapReduce-based spatial join; Hadoop; Optimized linear feature matching; Conflation

In the age of Big Data, there are many opportunities to collect and combine large volumes of geospatial data from a variety of agencies and from different volunteered geographic information (VGI) sources to facilitate scientific research and decision-making. However, two issues should be considered because of the efficiency and quality concerns. Firstly, the mining, harvesting, and geoprocessing of Big Geo-Data are very computationally intensive, thus one question is how to integrate high performance solutions with the support of CyberGIS to achieve the goals. Secondly, the problems of heterogeneity and incompatibility in geospatial data affect the conflation process, thus another question is how to optimize the feature matching procedure, which is one of the most challenging components in conflation.

To address the above issues, the cyberinfrastructure should have the scaling capability of storing, integrating, processing and visualizing Big Geo-data and have an easy-to-use, configurable user-interface to submit the processing jobs and to monitor the status of the computing platform. The Hadoop ecosystem is an ideal choice, since it provides a distributed file system (HDFS) and a scalable computation framework (MapReduce) by partitioning computation processes across connected host servers which are not necessary single-core-high-performance computers in a Hadoop cluster, but it lacks spatial analysis functionality. To this end, in this research, we build a spatially-enabled distributed platform based on Apache Hadoop with Cloudera Manager Packages, and integrate the Esri Geometry APIs to enable Hadoop cluster for scalable spatial analysis of geospatial data. Then, we set up a high-performance geoprocessing workflow for mining, harvesting, and analyzing various sources of VGI such as

Flickr geotagged photos, Twitter geotagged tweets, and Foursquare check-ins. Using examples of MapReduce-based spatial join, we demonstrate the high performance of this cyberinfrastructure in fast extracting, analyzing, and aggregating different feature types of place of interests (POI) at multiple geographic scales. Although post-processing is still necessary for a better quality output, this work offers new insights on connecting GIS tools to cloud computing environment for the next frontier of Big Geo-Data analysis.

In addition, we have been developing an application of Hadoop to parallelize an original algorithm of optimized linear feature matching for geospatial data conflation proposed by Li and Goodchild (2011) (*An optimisation model for linear feature matching in geographical data conflation. International Journal of Image and Data Fusion 2(4): 309–328*). We would discuss the strategies and challenges in speeding-up the algorithm. One impediment of this optimized conflation algorithm is that it requires much time to calculate the directed-Hausdorff-distance matrix for linear feature matching. Using the divide and conquer strategy, we can split the target features into different keys, i.e. the unified identifier for each linear feature, and then calculate the directed-Hausdorff-distance matrix in a parallel-sweeping manner on HDFS to improve the conflation efficiency. This procedure might be beneficial to conflating large volume of linear feature datasets provided by different government agencies such as US Census TIGER/Line and US Department of Transportation NTAD, or geospatial data downloaded from VGI website such as OpenStreetMap.