RINX 2.0: A Containerized Climate Raster Information Extraction System on OpenShift Cloud Environment

Jeff Blossom, Devika Jain, Jack Hayes, Heike Gibson, Sheryl Rifas-Shimann and Diane Gold

ABSTRACT:

RINX 2.0 (Raster INformation eXtraction) is an end-to-end solution developed by the authors for automatic extraction of information from large rasters datasets. The input for RINX is a set of geo-referenced raster datasets and a set of point locations from which the information is to be extracted. The output is a structured representation of extracted information from the raster datasets for each data point in CSV text format. The loading and processing of the input datasets to RINX 2.0 is accomplished using a combination of Bash and SQL scripts deployed in a containerized environment optimized for efficient automation. This environment uses the open technologies OpenShift and Crunchy Data to feed the input raster and point location data into the spatial database PostGIS for extraction. RINX 2.0 was created to aid the study of environmental conditions and how they affect the health of people over their lifespans for the Environmental influences on Child Health Outcomes project. The Environmental influences on Child Health Outcomes (ECHO) program is a nation-wide project in the United States funded by the National Institutes of Health. ECHO includes over 60 cohorts of children and their mothers and is aimed to help better understand effects of environmental exposures on child health and development. Daily meteorological and long-term climate conditions have been shown to have an adverse effect on health and are thus one of the environmental exposures of interest to the investigators in the ECHO program. One of the ECHO cohorts is Project Viva, a Boston, MA based longitudinal study including a cohort of some 2,000 mothers and children. For the Project Viva cohort we used RINX 2.0 to extract 7 daily climate variables (min., mean, max. and dew point temperature, precipitation, and min. and max. vapor pressure deficit) from the PRISM data for 5,219 address locations spanning variable time periods between from 1999 - 2023, for a total of 18,131,095 patient days. This produced a total of 126,917,665 climate observations, output into .csv format. Further, the mean and dew point temperatures were used to calculate relative humidity and absolute humidity for each day, producing an additional 36,262,190 observations for a total of 163,179,855 observations. This data is in the process of being combined with health outcome data, to analyze the effect climate may have on lung function and other systems. The entire process took 2 hours to load the rasters, and 1.5 days to calculate the 163M observations, and the architecture simplified the management and scaling of our project. It is estimated that traditional methods such as ArcGIS, QGIS, and R would have taken 2 months or more to extract the same amount of observations, thus demonstrating that RINX 2.0 saves considerable time and cost. With RINX 2.0 it will now be possible to rapidly and efficiently enrich additional ECHO, and other cohort address locations with climate exposure data.

KEYWORDS: *climate data, Big Data, cloud computing, raster data, geospatial, OpenShift, containers*

Jeff Blossom, Center for Geographic Analysis, Harvard University, Cambridge MA, USA

Devika Jain, Center for Geographic Analysis, Harvard University, Cambridge MA, USA

Jack Hayes, Center for Geographic Analysis, Harvard University, Cambridge MA, USA

Heike Gibson, Harvard T.H. Chan School of Public Health, Harvard University, Boston MA, USA

Sheryl Rifas-Shimann, Department of Population Medicine, Harvard Medical School, Boston, USA

Diane Gold, Harvard T.H. Chan School of Public Health, Harvard University, Boston MA, USA