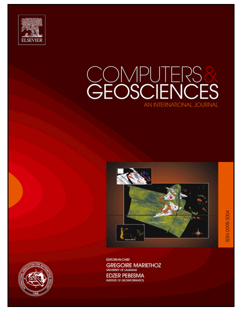
Accepted Manuscript



ClimateSpark: An in-memory distributed computing framework for big climate data analytics

Fei Hu, Chaowei Yang, John L. Schnase, Daniel Q. Duffy, Mengchao Xu, Michael K. Bowen, Tsengdar Lee, Weiwei Song

PII: S0098-3004(16)30609-4

DOI: 10.1016/j.cageo.2018.03.011

Reference: CAGEO 4110

To appear in: *Computers and Geosciences*

Received Date: 30 October 2016

Revised Date: 3 December 2017

Accepted Date: 19 March 2018

Please cite this article as: Hu, F., Yang, C., Schnase, J.L., Duffy, D.Q., Xu, M., Bowen, M.K., Lee, T., Song, W., ClimateSpark: An in-memory distributed computing framework for big climate data analytics, *Computers and Geosciences* (2018), doi: 10.1016/j.cageo.2018.03.011.

ClimateSpark: An In-memory Distributed Computing Framework for Big Climate Data Analytics

Fei Hu[1], Chaowei Yang[1*], John L. Schnase[2], Daniel Q. Duffy[2], Mengchao Xu[1], Michael K. Bowen[2], Tsengdar Lee[3], Weiwei Song[1]

[1]NSF Spatiotemporal Innovation Center, George Mason University, Fairfax, VA, 22030-4444

[2]NASA Goddard Space Flight Center, Greenbelt, MD, 20771

[3]NASA Headquarters, Washington DC, DC

*Corresponding Author: cyang3@gmu.edu, 7039934742

Abstract: The unprecedented growth of climate data creates new opportunities for climate studies, and yet big climate data pose a grand challenge to climatologists to efficiently manage and analyze big data. The complexity of climate data content and analytical algorithms increases the difficulty of implementing algorithms on high performance computing systems. This paper proposes an in-memory, distributed computing framework, *ClimateSpark*, to facilitate complex big data analytics and time-consuming computational tasks. Chunking data structure improves parallel I/O efficiency, while a spatiotemporal index is built for the chunks to avoid unnecessary data reading and preprocessing. An integrated, multi-dimensional, array-based data model (ClimateRDD) and ETL operations are developed to address big climate data variety by integrating the processing components of the climate data lifecycle. ClimateSpark utilizes Spark SQL and Apache Zeppelin to develop a web portal to facilitate the interaction among climatologists, climate data, analytic operations and computing resources (e.g., using SQL query and Scala/Python notebook). Experimental results show that ClimateSpark conducts different spatiotemporal data queries/analytics with high efficiency and data locality. ClimateSpark is easily adaptable to other big multiple-dimensional, array-based datasets in various geoscience domains.

Keywords: Big Data, high performance computing, array-based data model, climate data analytics, Apache Spark, Geospatial Cyberinfrastructure, Cloud Computing.

1. Introduction

Climate science is a big data domain with unprecedented growth of climate data (Schenase et al., 2014; Yang et al., 2017a). Climate scientists analyze past observations, integrate billions of daily earth observations and perform climate-change simulations, all of which produce large volumes of data (Edwards, 2010; Yang et al., 2017b). National Aeronautics and Space Administration (NASA) projected the size of the climate change data repositories to grow to 350 petabytes by 2030 (Skyland, 2012). The United Nation's Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5) was based on