Computers, Environment and Urban Systems 62 (2017) 210–221

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

A high performance query analytical framework for supporting data-intensive climate studies

Zhenlong Li [a,*], Qunying Huang [b], Gregory J. Carbone [a], Fei Hu [c]

[a] Department of Geography, University of South Carolina, Columbia, SC 29208, United States
[b] Department of Geography, University of Wisconsin-Madison, Madison, WI 53706, United States
[c] Spatiotemporal Innovation Center, George Mason University, Fairfax, VA 22030, United States

ARTICLE INFO

Article history:
Received 15 December 2015
Received in revised form 29 November 2016
Accepted 6 December 2016
Available online xxxx

Keywords:
Big data
Cyberinfrastructure
Climate analysis
Query analytics
Cloud computing
Hive

ABSTRACT

Climate observations and model simulations produce vast amounts of data. The unprecedented data volume and the complexity of geospatial statistics and analysis requires efficient analysis of big climate data to investigate global problems such as climate change, natural disasters, diseases, and other environmental issues. This paper introduces a high performance query analytical framework to tackle these challenges by leveraging Hive and cloud computing technologies. With this framework, we propose grid transformation, a new perspective for complex climate analysis that applies a series of atomic transformations to terabytes of climate data using SQL-style query (HiveQL). Specifically, we introduce four types of grid transformations (temporal, spatial, local, and arithmetic) to support a broad range of climate analyses, from the basic spatiotemporal aggregation to more sophisticated anomaly detection. Each query is processed as MapReduce tasks in a highly scalable Hadoop cluster as the parallel processing engine. Big climate data are directly stored and managed in a Hadoop Distributed File System without any data format conversion. A prototype is developed to evaluate the feasibility and performance of the framework. Experimental results show that complex and data-intensive climate analysis can be conducted using intuitive SQL queries with good flexibility and performance. This research provides a building block and practical insights in establishing a cyberinfrastructure that provides a high performance and collaborative environment for data-intensive geospatial applications in climate science.

1. Introduction

Global climate models simulate the earth-atmospheric-ocean system and have been used to project future climate changes. In building a better understanding of the climate, environment, and anthropogenic influences on the system, these models create massive amounts of data. For example, the data volume provided for the Fifth Assessment Report (AR5) of the Intergovernmental Panel on Climate Change (IPCC) is 1.7 petabytes (WDCC, 2015). Additionally, recent advancements in data acquisition technology, including remote sensing, have allowed us to collect massive volumes of observational data. Efficient analysis of vast climate data sets enables scientists to address fundamental questions about global and regional climate, such as trends, anomalies, and extremes (Das & Parthasarathy, 2009).

These climate data play a critical role in understanding how the complex climate system works. They also typify the characteristics of "Big Data" (Manovich, 2011) and pose several grand challenges to data management, processing, and analytical infrastructure in three aspects:

- **Volume**: Climate data are massive in volume. NASA Goddard Institute for Space Studies (GISS) ModelE, for example, produces 2.5 terabytes of data with one ensemble run (Sun et al., 2012). The Data Active Archive Centers (DAACs) provide tens of terabytes of high quality NASA data products for inferring a large number of physical phenomena from Earth observing instrument records (Chapman, Simon, Nguyen, & Halem, 2013). Such magnitude exceeds the capability of any data storage and processing system on a stand-alone computer (Wang et al., 2013) because the computational requirements involved go beyond what a typical workstation can support. Such big climate data calls for distributed storage media and innovative computing architecture.
- **Velocity**: Climate data from real-time monitoring, Earth observation systems (EOS), or model simulations are generated dynamically and continuously. With the advancement of satellite and sensor systems, Earth observation and monitoring data are collected and accumulated at an unprecedented speed. These real-time data, while essential to understanding Earth-system processes, require an adaptable system to dynamically adjust to different acquisition rates coming from

* Corresponding author.
E-mail addresses: zhenlong@sc.edu (Z. Li), qhuang46@wisc.edu (Q. Huang), carbone@mailbox.sc.edu (G.J. Carbone), fhu@gmu.edu (F. Hu).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.12.003
0198-9715/© 2016 Elsevier Ltd. All rights reserved.

different sensors. Meanwhile, climate models can be run on thousands of remote servers with different parameterizations (Stainforth et al., 2002) to verify and validate a climate model or test our hypothesis of climate change for better decision making (Li et al., 2014). The output from such a distributed network is uploaded to the data server concurrently and continuously, generating unpredictable traffic loads. Traditional climate data handling system cannot adjust to such dynamic processes because they lack scalable computing infrastructures to handle streaming climate data. Consequently, climate data processing must be extended to cope with this dynamic information flow.

- **Variety**: Climate data are complex in nature, and often produced and stored in different forms. For example, many different data models, formats, standards, tools, services, and terminology are defined and designed to organize and store EOS data by geoscience communities. Additionally, climate data are highly dimensional, including spatio-temporal dimensions and a variety of climate attributes or variables, such as temperature, precipitation, pressure and relative humidity. As a result, analytic operations over climate data need to quickly retrieve minimum inputs that indicate the climate variable, spatial extent, and temporal scope (Schnase, Duffy, McInerney, Webster, & Lee, 2015). To leverage the widely available data sources for studying climate processes and mining hidden patterns, the community urgently needs a flexible data structure and a storage and analytical framework that can easily organize and process multi-sourced, high-dimensional climate data.

Addressing these challenges, known as 3Vs, requires efficient data management strategies, complex parallel algorithms, and scalable computing resources. Currently, NoSQL systems and large-scale data platforms based on MapReduce paradigm, such as Hadoop, are widely used for big data management and analytics (Witayangkurn, Horanont, & Shibasaki, 2013). However, current implementations of MapReduce-based management and data processing systems do not offer enough support for spatiotemporal data query and analytics. Array-based database systems, such as RasDaMan (Baumann, Dehmel, Furtado, Ritsch, & Widmann, 1998, 1999) and SciDB (Brown, 2010; Cudré-Mauroux et al., 2009), have emerged as scalable database solutions to store and retrieve massive, multi-dimensional datasets that were traditionally stored and organized in various formats, such as NetCDF, and HDF. However, these solutions have proved to be limited in handling massive, multi-dimensional big data because they lack high-performance support, and scalability (Liu, 2014). Also, these systems require conversion of large data sets to specific formats – an undesirable step considering the volume.

To bridge the gap and advance data-intensive climate studies, this paper introduces a high-performance query analytical framework, which allows us to perform data-intensive analysis in parallel using intuitive SQL-style queries. The remaining paper is organized as follows: Section 2 reviews the current status for big climate data analytics and query processing techniques; Sections 3 & 4 detail the framework and methodologies; Section 5 evaluates the framework with a proof-of-concept prototype; and Section 6 offers summary remarks.

2. Related work

2.1. Evolution of climate data analytics

In the past decade, a variety of analytical tools dealing with climate data has been developed to process and analyze climate data. For example, NASA developed a cross-platform tool, known as Panoply,[1] to plot and explore geo-referenced and other arrays from NetCDF, HDF, GRIB, and other datasets. Chapman et al. (2013) described a statistical aggregation engine "Gridderama" for climate trend analysis. ViSUS/Climate Data and Analysis Tools (CDAT) system was implemented to provide the user with a variety of visualization techniques for data exploration and analysis (Potter et al., 2009). Based on CDAT, the Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT) was developed to enable parallel visualization and analysis of computer model output resulting from high-resolution, long-term, climate change projections (Santos et al., 2013). UV-CDAT consists of several components, including CDAT (used in the climate community since 1995), VisTrails, DV3D, ParaView, and the Visualization Toolkit (VTK, an open source, object-oriented library for visualization and analysis). Despite strong analytical capabilities, these systems have notable limitations. Users need to download and install multiple packages on their machines, and to identify and manually download the data. Additionally, they are not flexible, and the functionalities are static.

With the advancement of Internet and computing infrastructure, climate data analytical systems are transitioning from the traditional models – running on stand-alone desktop computers or workstations – to the Web. Currently, many online systems are developed to deliver analytics capabilities of visualizing, analyzing, and accessing vast amounts of climate data (e.g., Acker & Leptoukh, 2007; Sun et al., 2012; Luo et al., 2015). These online systems overcome the limitations of the stand-alone systems by managing large volume datasets over the Internet and support interactive data analysis.

Cloud computing emerges as a new computing paradigm with a flexible stack of massive computing, storage, and software services to support big data analytics in a scalable manner at low cost (e.g. Yang et al., 2011; Li, Yang, Yu, Liu, & Sun, 2015; Li Hodgson & Li, 2016). Various climate applications are leveraging such flexible and on-demand cloud resources, for example, running coupled atmosphere-ocean climate models (Evangelinos & Hill, 2008), and supporting dust storm forecasting (Huang et al., 2013). Li et al. (2014) proposed Model as a Service (MaaS), a cloud-computing solution for Internet-based climate model simulation for the public. With Analytics as a Service (AaaS), cloud computing not only provides the computing infrastructure to support Big Data handling, but also provides a computing model to support to discovery in Big Data (Lomotey & Deters, 2014). Specifically, the concept of Climate AaaS (CAaaS) emerged to provide climate data analytics as web services in a high performance and scalable way (Schnase et al., 2014; Schnase et al., 2015). This paper contributes to the literature another solution for conducting large-scale climate data analytics using the SQL-style query in a scalable, flexible, and high-performance manner.

2.2. High performance computing for handling raster data

Efficient processing and analysis of massive spatial raster data (e.g. climate data in HDF format) is paramount to geospatial problem-solving and decision-making. Big raster data derived from imaging and spatial applications require high performance computing support and the development of novel techniques. Currently, various HPC-enabled techniques are proposed to process massive raster data. Zhang, You, and Gruenwald (2010), for example, introduced an approach for fast indexing of large-scale raster geospatial data using Graphics Processing Unit (GPU) computing which improved speed by twenty-three times in controlled experiments. More similar works on spatial big data processing (e.g., spatial indexing and spatial joins) on GPUs and GPU-accelerated clusters, were reported in a later study (Zhang, You, & Gruenwald, 2015). Scott, Backus, and Anderson (2014) also developed a GPU cluster framework to process geospatial raster data in parallel across tiles to extract refined geospatial information.

MapReduce frameworks, e.g. Hadoop, can easily scale data computation over multiple computing nodes and therefore are well suited for spatial data processing and spatial analysis in a variety of geospatial applications, such as satellite data processing and analysis (Golpayegani & Halem, 2009; Li, Yang, Liu, Hu, & Jin, 2016). Examples of remote sensing data handling based on the MapReduce framework include Sobel

[1] http://www.giss.nasa.gov/tools/panoply/