Computers, Environment and Urban Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service

John L. Schnase [a,*], Daniel Q. Duffy [b], Glenn S. Tamkin [a], Denis Nadeau [a], John H. Thompson [b], Cristina M. Grieg [c], Mark A. McInerney [a], William P. Webster [a]

[a] Office of Computational and Information Sciences and Technology, NASA Goddard Space Flight Center, Greenbelt, MD 20771, United States
[b] NASA Center for Climate Simulation, NASA Goddard Space Flight Center, Greenbelt, MD 20771, United States
[c] Department of Computational Data Sciences, George Mason University, Fairfax, VA 22030, United States

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
MapReduce
Hadoop
Data analytics
Data services
Cloud Computing
Generativity
iRODS
MERRA
ESGF
BAER

ABSTRACT

Climate science is a Big Data domain that is experiencing unprecedented growth. In our efforts to address the Big Data challenges of climate science, we are moving toward a notion of Climate Analytics-as-a-Service (CAaaS). We focus on analytics, because it is the knowledge gained from our interactions with Big Data that ultimately produce societal benefits. We focus on CAaaS because we believe it provides a useful way of thinking about the problem: a specialization of the concept of business process-as-a-service, which is an evolving extension of IaaS, PaaS, and SaaS enabled by Cloud Computing. Within this framework, Cloud Computing plays an important role; however, we see it as only one element in a constellation of capabilities that are essential to delivering climate analytics as a service. These elements are essential because in the aggregate they lead to generativity, a capacity for self-assembly that we feel is the key to solving many of the Big Data challenges in this domain. MERRA Analytic Services (MERRA/AS) is an example of cloud-enabled CAaaS built on this principle. MERRA/AS enables MapReduce analytics over NASA's Modern-Era Retrospective Analysis for Research and Applications (MERRA) data collection. The MERRA reanalysis integrates observational data with numerical models to produce a global temporally and spatially consistent synthesis of 26 key climate variables. It represents a type of data product that is of growing importance to scientists doing climate change research and a wide range of decision support applications. MERRA/AS brings together the following generative elements in a full, end-to-end demonstration of CAaaS capabilities: (1) high-performance, data proximal analytics, (2) scalable data management, (3) software appliance virtualization, (4) adaptive analytics, and (5) a domain-harmonized API. The effectiveness of MERRA/AS has been demonstrated in several applications. In our experience, Cloud Computing lowers the barriers and risk to organizational change, fosters innovation and experimentation, facilitates technology transfer, and provides the agility required to meet our customers' increasing and changing needs. Cloud Computing is providing a new tier in the data services stack that helps connect earthbound, enterprise-level data and computational resources to new customers and new mobility-driven applications and modes of work. For climate science, Cloud Computing's capacity to engage communities in the construction of new capabilities is perhaps the most important link between Cloud Computing and Big Data.

Published by Elsevier Ltd.

1. Introduction

The term "Big Data" is used to describe data sets that are too large and complex to be worked with using commonly-available tools (Snijders, Matzat, & Reips, 2012). Climate science represents a Big Data domain that is experiencing unprecedented growth (Edwards, 2010). NASA's climate change repositories alone are projected to grow to 350 petabytes by 2013 (Skytland, 2012). Some of the major Big Data challenges facing climate science are easy to understand: large repositories mean that the data sets themselves cannot be moved: instead, analytical operations need to migrate to where the data reside; complex analyses over large repositories requires high-performance computing; large amounts of information increases the importance of metadata, provenance management, and discovery; migrating codes and analytic products within a growing network of storage and computational resources creates a need for fast networks, intermediation, and resource balancing; and, importantly, the ability to respond quickly to customer demands for new and often unanticipated uses for climate data requires greater agility in building and deploying applications. It is useful to situate the Big Data challenges of the climate domain in this larger context, because doing so helps us understand where innovation can yield improvements.

* Corresponding author. Tel.: +1 301 286 4351.
E-mail address: John.L.Schnase@NASA.gov (J.L. Schnase).

0198-9715/$ - see front matter Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.compenvurbsys.2013.12.003

Please cite this article in press as: Schnase, J. L., et al. MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service. *Computers, Environment and Urban Systems* (2014), http://dx.doi.org/10.1016/j.compenvurbsys.2013.12.003

Cloud Computing is one of several technologies often invoked as a solution to Big Data challenges. However, the technical definition of "Cloud Computing" is so variously interpreted that the term has become jargonized (Mell & Grace, 2011). That Cloud Computing is both ubiquitous and ambiguous points to the need to examine carefully how Cloud Computing enables.

1.1. Climate Analytics-as-a-Service (CAaaS)

In our efforts to address the Big Data challenges of climate science, we are moving toward a notion of Climate Analytics-as-a-Service (CAaaS). We focus on analytics, because it is the knowledge gained from our interactions with Big Data that ultimately produce societal benefits. We focus on CAaaS because we believe it provides a useful way of thinking about the problem: a specialization of the concept of business process-as-a-service, which is an evolving extension of IaaS, PaaS, and SaaS enabled by Cloud Computing. Within this framework, Cloud Computing plays an important role; however, we see it as only one element in a constellation of capabilities that are essential to delivering climate analytics as a service. These elements are essential because in the aggregate they lead to *generativity* – a capacity for self-assembly that we feel is the key to solving many of the Big Data challenges in this domain.

1.2. Generative technologies

Generativity refers to a system's capacity to produce unanticipated change through unfiltered contributions from broad and varied audiences (Zittrain, 2008). The concept highlights aspects of an innovation or process that enable an autocatalytic feeding-forward that can help make growth, further innovation, and success possible. Generativity connects inputs from diverse people and groups, who may or may not be working in concert, with emergent and unanticipated outputs. How much the system facilitates participant contribution is a function of both *technological design* and *social behavior* (Baker & Bowker, 2007). A system's generativity describes not only its objective characteristics, but also the ways the system relates to its users and the ways users relate to one another. In turn, these relationships reflect how much the users identify as contributors or participants, rather than as mere consumers.

The Internet itself, modern operating systems, Apple's iTunes, Twitter, Facebook, and the emerging infrastructure for mobile application development are examples of generative systems. In both cases, the design elements contributing to their generative potential are easy to see. The Internet's framers made *simplicity* a core value, defining in the process the classic end-to-end argument that most features in a network should be implemented at its computer endpoints rather than by the network itself, which appropriately implements only those functions that are universally useful (Saltzer, Reed, & Clark, 1984). To do otherwise might have tilted the generic network toward specific uses and limited its potential for growth. (Consider, for example, the proprietary, non-generative, and now defunct CompuServe network.).

Zittrain (2008) identifies five properties of generative systems:

(1) *How extensively a system or technology leverages a set of possible tasks:* Leverage makes a difficult job easier, and, in general, the more a system can do, the more capable it is of producing change.
(2) *How well it can be adapted to a range of tasks:* Adaptability enables new, unintended, and innovative uses of a technology. It broadens the technology's use.
(3) *How easily new contributors can master it:* Ease of Mastery reflects how easy it is for broad audiences to understand how to adopt and adapt it. The more useful a technology is both to the neophyte and the expert, the more generative it is.
(4) *How accessible it is to those ready and able to build on it:* Accessibility makes it easier to obtain the technology and the information necessary to achieve mastery. The more accessible, the more generative.
(5) *How transferable any changes are to others, including non-experts:* Transferability reflects how easily changes in the technology can be conveyed to others.

A major deficiency in any one factor greatly reduces overall generativity. Conversely, the more these five qualities are maximized, the easier it is for a system to welcome contributions from outsiders as well as insiders. In general, generative tools are more basic and less specialized for accomplishing a particular purpose.

In the remainder of this paper, we illustrate how we are translating these concepts into reality: we describe the context in which we are working; the technology foundations important to us, including our definition and rationale for the generative elements we feel are crucial; a specific project, MERRA Analytic Services, and applications that demonstrate these capabilities in action; ways that Cloud Computing are contributing to the effort; and, finally, our plans for the future.

2. Background – The NASA Center for Climate Simulation and climate science as a Big Data domain

Our understanding of the Earth's processes is based on a combination of observational data records and mathematical models. The size of NASA's space-based observational data sets is growing dramatically as new missions come online. However, a potentially bigger data challenge is posed by the work of climate scientists, whose models are regularly producing data sets of hundreds of terabytes or more (Edwards, 2010; Webster, 2013).

The NASA Center for Climate Simulation (NCCS) provides state-of-the-art supercomputing and data services specifically designed for weather and climate research (NCCS, 2013). The NCCS maintains advanced data capabilities and facilities that allow researchers within and beyond NASA to create and access the enormous volume of data generated by weather and climate models. Tackling the problems of data intensive science is an inherent part of the NCCS mission.

There are two major challenges posed by the data intensive nature of climate science. There is the need to provide complete life-cycle management of large-scale scientific repositories. This capability is the foundation upon which a variety of data services can be provided, from supporting active research to large-scale data federation, data publication and distribution, and archival storage (Berman, 2008). We think of this aspect of our mission as climate data services.

The other data intensive challenge has to do with how these large datasets are used: data analytics – the capacity to perform useful scientific analyses over enormous quantities of data in reasonable amounts of time. In many respects this is the biggest challenge; without effective means for transforming large scientific data collections into meaningful scientific knowledge, our mission fails. It is against this backdrop that the NCCS began looking at CAaaS as a potential element in our technological and organizational response to changing demands.

3. Technology foundations – Toward a generative ecology for Climate Analytics-as-a-Service

We believe there are five essential technology elements that contribute to building a generative context for Climate Analytics-as-a-Service: high-performance, data-proximal analytics; integrative data management; software appliance virtualization; adaptive

Please cite this article in press as: Schnase, J. L., et al. MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service. *Computers, Environment and Urban Systems* (2014), http://dx.doi.org/10.1016/j.compenvurbsys.2013.12.003

Download English Version:

https://daneshyari.com/en/article/4965243

Download Persian Version:

https://daneshyari.com/article/4965243

Daneshyari.com