Computers, Environment and Urban Systems 61 (2017) 120–128

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Utilizing Cloud Computing to address big geospatial data challenges

Chaowei Yang *, Manzhu Yu, Fei Hu, Yongyao Jiang, Yun Li

NSF Spatiotemporal Innovation Center, George Mason University, Fairfax, VA 22030, United States

ARTICLE INFO

Article history:
Received 7 December 2015
Received in revised form 20 October 2016
Accepted 20 October 2016
Available online 2 November 2016

Keywords:
Big Data
Cloud Computing
Spatiotemporal data
Geospatial science
Smart cities

ABSTRACT

Big Data has emerged with new opportunities for research, development, innovation and business. It is characterized by the so-called four Vs: volume, velocity, veracity and variety and may bring significant value through the processing of Big Data. The transformation of Big Data's 4 Vs into the 5th (value) is a grand challenge for processing capacity. Cloud Computing has emerged as a new paradigm to provide computing as a utility service for addressing different processing needs with a) on demand services, b) pooled resources, c) elasticity, d) broad band access and e) measured services. The utility of delivering computing capability fosters a potential solution for the transformation of Big Data's 4 Vs into the 5th (value). This paper investigates how Cloud Computing can be utilized to address Big Data challenges to enable such transformation. We introduce and review four geospatial scientific examples, including climate studies, geospatial knowledge mining, land cover simulation, and dust storm modelling. The method is presented in a tabular framework as a guidance to leverage Cloud Computing for Big Data solutions. It is demostrated throught the four examples that the framework method supports the life cycle of Big Data processing, including management, access, mining analytics, simulation and forecasting. This tabular framework can also be referred as a guidance to develop potential solutions for other big geospatial data challenges and initiatives, such as smart cities.

1. Introduction

Earth observation and model simulation produce tera- to peta- bytes of data daily (Yang, Raskin, Goodchild, and Gahegan, 2010). Non-traditional, geospatial data acquisition methods, such as social media (Romero, Galuba, Asur, and Huberman, 2011), phone conversations (Frias-Martinez, Virseda, Rubio, and Frias-Martinez, 2010) and unmanned aerial vehicles (Einav and Levin, 2013), produce geospatial data at even faster speeds. In addition to the large Volume (Marr, 2015; Hsu, Slagter, and Chung, 2015), geospatial data exist in a Variety of forms and formats for different applications, their accuracy and uncertainty span across a wide range as defined by Veracity, and data are produced in a fast Velocity through real time sensors (Fig. 1). With unprecedented information and knowledge embedded, these big geospatial data can be processed for adding Value to better scientific research, engineering development and business decisions (Lee and Kang, 2015). They are envisioned to provide innovation and advancements to improve our lives and understanding of the Earth systems (Mayer-Schönberger and Cukier, 2013) when transformed from the first four Vs to the last V (value) through advancements in a variety of geospatial domains (Fig. 1).

Such transformations pose grand challenges to data management and access, analytics, mining, system architecture and simulations (Yang, Huang, Li, Liu, and Hu, 2016). For example, the first challenge is how to deal with the Variety and Veracity of Big Data to produce a fused dataset that can be utilized in a single decision support system (Kim, Trimi, and Chung, 2014). Another issue is how to deal with the velocity of Big Data to have scalable and extensible processing power based on the fluctuation of the data feed (Ammn and Irfanuddin, 2013). Supporting on-demand or timely data analytical functionalities also pose significant challenges for creating the Value (Fan and Liu, 2013; Chen and Zhang, 2014; Jagadish et al., 2014).

Cloud Computing has emerged as a new paradigm to provide computing as a utility service with five advantageous characteristics (Fig. 1 bottom two layers): a) rapid and elastic provisioning computing power; b) pooled computing power to better utilize and share resources; c) broadband access for fast communication; d) on demand access for computing as utility services; and e) pay-as-you-go for the parts used without a significant upfront cost like that of traditional computing resources (Yang, Xu, and Nebert, 2013). Service-oriented architecture is adopted in Cloud Computing and enables "everything as a service", including Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) (Mell and Grance, 2011). While redefining the possibilities of geospatial science and Digital Earth (Yang et al., 2013), Cloud Computing engaging Big Data enlightens potential solutions for big geospatial data problems in various geosciences and relevant domains.

However, utilizing Cloud Computing to address Big Data issues is still in its infancy, and it is a daunting task on how the five advantageous

* Corresponding author.
E-mail address: cyang3@gmu.edu (C. Yang).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.010
0198-9715/

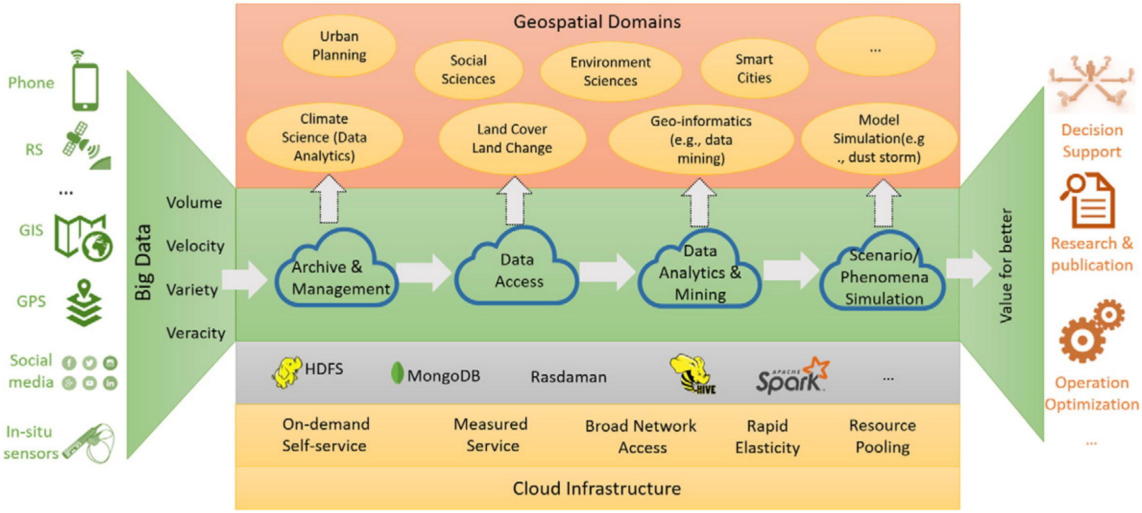

Fig. 1. Cloud Computing provides critical supports to the processing of Big Data to address the 4Vs to obtain Value for better decision support, research, and operations for various geospatial domains.

characteristics can address the first four Vs of Big Data to reach the 5th V (Fig. 1). This paper illustrates how Cloud Computing supports the transformation with four scientific examples including climate studies, knowledge mining, land-use and land cover change analysis, and dust storm simulation. These four examples are highly representative and can be easily adopted to other environmental and urban research fields, such as smart cities (Batty, 2013; Mitton, Papavassiliou, Puliafito, and Trivedi, 2012; Odendaal, 2003). The big geospatial data life cycle (data management and access, analyses/mining, phenomena/scenario simulation) are examined through the four examples and detailed in each example section (Table 1). For example, 2.1 is filled in the intersection cell of on-demand self-service and volume of Table 1. This means that 2.1 details how the volume (of big climate data) are addressed with the on-demand self-service of Cloud Computing.

2. Utilizing Cloud Computing to support climate analytics

The interrelated climate changes, such as greater incidence of heavy downpours and increased riverine flooding, are increasingly compromising urban infrastructure (Rosenzweig, Solecki, Hammer, and Mehrotra, 2011). Meanwhile human activities (e.g. the burning of fossil fuels) heavily impacted the global environment in the past 50 years (Bulkeley and Betsill, 2005). In order to understand climate change and its impacts to environmental and urban issues, the big climate data observed in the past and simulated for the future should be well managed and analyzed. However, both observation and simulation produce Big Data. For example, the next IPCC report will be based on 100 + petabytes of data, and NASA will produce 300 + petabytes of climate data by 2030 (Skytland, 2012). These data differ in format, spatiotemporal resolution, and study objective (Schnase et al., 2014). Big Data can help advance the understanding of climate phenomena and help identify how impacts of climate change on society and ecosystems can be remedied, such as detecting global temperature anomalies and investigating spatiotemporal distribution of extreme weather events, especially over highly populated regions (such as urban areas, Das and Parthasarathy, 2009; Debbage and Shepherd, 2015).

Table 1
The Big Data challenges as illustrated in the four examples are addressed by relevant cloud advantages to reach the Big Data Value and achieve the research, engineering and application objectives.

	On-demand Self-service	Broad network access	Resource pooling	Rapid elasticity	Measured service
Volume	2.1	4.1	2.1	2.1, 3.1, 3.2, 4.1, 4.2, 4.3, 5.1	4.1
Veracity	2.1	3.1, 5.3			
Velocity			2.1	4.1	4.3
Variety		3.1, 5.2	2.1		
Value	2.1, 3.2			2.1	3.2

There are several challenges in the use of Big Data: a) the volume and velocity of big climate data have far exceeded the stand-alone computer's storage and computing ability; b) the variety of climate data in format and spatiotemporal resolution make it difficult to find an easy-to-use tool to analyze climate data; c) the veracity in model simulation is a concern for climate scientists of the uncertainties and mixed model qualities (Murphy et al., 2004). The combined complexities of volume, velocity, variety, and veracity can be addressed with cloud-based, advanced data management strategies and a service-oriented data analytical architecture to help process, analyze and mine climate data.

2.1. Advanced spatiotemporal index for big climate data management

The hundreds of petabytes of climate data can only be managed in a distributed and scalable environment. Cloud Computing could help the management as follows: a) provisioning on-demand flexible virtual machines (VM) according to the volume of climate data; and b) automatically deploying HDFS, Hadoop Distributed File System, on the VMs to build a distributed filesystem. Data can be maintained in native format instead of sequenced text for saving storage space. A logical data architecture is also built to facilitate fast identification, access, and analyses (Li, Hu et al., 2016; Li, Yang et al., 2016). The core architecture is a spatiotemporal index (Li, Hu et al., 2016; Li, Yang et al., 2016) for the multi-dimensional climate data stored on HDFS. The index maps data content onto the byte, file and node levels within the HDFS. Nine components are used for the index and include: space, time and shape information describe the data grid's logical information which correlates to data query, byte offset, byte length, compression code, node list and file path identify specific location on the HDFS. This index enables users to directly locate and access data with exact spatiotemporal and content description.

In details, the space and time attributes in the spatiotemporal index will identify the grids overlapped with a spatiotemporal bounding box. The node list attribute is leveraged to deliver the computing programs to the node where the grids are stored. Then the computing programs can read the data as a data stream with high data locality, according to the byte offset, byte length, and compression code attributes. The