Taylor & Francis
Taylor & Francis Group

International Journal of Digital Earth

ISSN: 1753-8947 (Print) 1753-8955 (Online) Journal homepage: http://www.tandfonline.com/loi/tjde20

Big Data and cloud computing: innovation opportunities and challenges

Chaowei Yang, Qunying Huang, Zhenlong Li, Kai Liu & Fei Hu

To cite this article: Chaowei Yang, Qunying Huang, Zhenlong Li, Kai Liu & Fei Hu (2017) Big Data and cloud computing: innovation opportunities and challenges, International Journal of Digital Earth, 10:1, 13-53, DOI: 10.1080/17538947.2016.1239771

To link to this article: http://dx.doi.org/10.1080/17538947.2016.1239771

Published online: 03 Nov 2016.

INTERNATIONAL JOURNAL OF DIGITAL EARTH, 2017
VOL. 10, NO. 1, 13–53
http://dx.doi.org/10.1080/17538947.2016.1239771

REVIEW ARTICLE

OPEN ACCESS

Big Data and cloud computing: innovation opportunities and challenges

Chaowei Yang[a], Qunying Huang[b], Zhenlong Li[c], Kai Liu[a] and Fei Hu[a]

[a]NSF Spatiotemporal Innovation Center, George Mason University, Fairfax, VA, USA; [b]Department of Geography, University of Wisconsin-Madison, Madison, WI, USA; [c]Department of Geography, University of South Carolina, Columbia, SC, USA

ABSTRACT

Big Data has emerged in the past few years as a new paradigm providing abundant data and opportunities to improve and/or enable research and decision-support applications with unprecedented value for digital earth applications including business, sciences and engineering. At the same time, Big Data presents challenges for digital earth to store, transport, process, mine and serve the data. Cloud computing provides fundamental support to address the challenges with shared computing resources including computing, storage, networking and analytical software; the application of these resources has fostered impressive Big Data advancements. This paper surveys the two frontiers – Big Data and cloud computing – and reviews the advantages and consequences of utilizing cloud computing to tackling Big Data in the digital earth and relevant science domains. From the aspects of a general introduction, sources, challenges, technology status and research opportunities, the following observations are offered: (i) cloud computing and Big Data enable science discoveries and application developments; (ii) cloud computing provides major solutions for Big Data; (iii) Big Data, spatiotemporal thinking and various application domains drive the advancement of cloud computing and relevant technologies with new requirements; (iv) intrinsic spatiotemporal principles of Big Data and geospatial sciences provide the source for finding technical and theoretical solutions to optimize cloud computing and processing Big Data; (v) open availability of Big Data and processing capability pose social challenges of geospatial significance and (vi) a weave of innovations is transforming Big Data into geospatial research, engineering and business values. This review introduces future innovations and a research agenda for cloud computing supporting the transformation of the volume, velocity, variety and veracity into values of Big Data for local to global digital earth science and applications.

ARTICLE HISTORY
Received 23 July 2016
Accepted 17 September 2016

KEYWORDS
Spatiotemporal computing; digital earth; geospatial cyberinfrastructure; geoinformatics; CyberGIS; Smart cities

1. Introduction

Big Data refers to the flood of digital data from many digital earth sources, including sensors, digitizers, scanners, numerical modeling, mobile phones, Internet, videos, e-mails and social networks. The data types include texts, geometries, images, videos, sounds and combinations of each. Such data

CONTACT Chaowei Yang ✉ cyang3@gmu.edu NSF Spatiotemporal Innovation Center, George Mason University, Fairfax, VA 22030, USA

can be directly or indirectly related to geospatial information (Berkovich and Liao 2012). The evolution of technologies and human understanding of the data have shifted data handling from the more traditional static mode to an accelerating data arena characterized by volume, velocity, variety, veracity and value (i.e. 5Vs of Big Data; Marr 2015). The first V refers to the volume of data which is growing explosively and extends beyond our capability of handling large data sets; volume is the most common descriptor of Big Data (e.g. Hsu, Slagter, and Chung 2015). Velocity refers to the fast generation and transmission of data across the Internet as exemplified by data collection from social networks, massive array of sensors from the micro (atomic) to the macro (global) level and data transmission from sensors to supercomputers and decision-makers. Variety refers to the diverse data forms and in which model and structural data are archived. Veracity refers to the diversity of quality, accuracy and trustworthiness of the data. All four Vs are important for reaching the 5th V, which focuses on specific research and decision-support applications that improve our lives, work and prosperity (Mayer-Schönberger and Cukier 2013).

The evolution of Big Data, especially its adoption by industry and government, expands the content/meaning of Big Data. The original volume-based definition now encompasses the data itself, relevant technologies and expertise to help generate, collect, store, manage, process, analyze, present and utilize data, as well as the information and knowledge derived. For example, the Big Earth Data Initiative (BEDI, The Whitehouse 2014; Ramapriyan 2015) designates Big Data as an investment opportunity and a 'calling card' for advancing earth science and digital earth using Big Data and relevant processing technologies. For the geospatial domain, Big Data has evolved along a path from purely data to a broader concept including data, technology and workforce. The focus is the geographic aspects of Big Data from Social, Earth Observation (EO), Sensor Observation Service (SOS), Cyber Infrastructure (CI), social media and business. For example, EO generates terabytes (TB) of images daily; climate simulations by the IPCC (Intergovernmental Panel on Climate Change) produce hundreds of peta-bytes (PB) for future climate analyses; and SOS produces even more from sensor web and citizen as sensors (Goodchild 2007). Social and business data are generated at a faster pace with specific geographic and temporal footprints (i.e. spatiotemporal data). Tien (2013) investigated the data quality and information to further improve innovation of Big Data to address the 14 engineering challenges identified by National Academy of Engineering in 2008 and advance the 10 breakthrough technologies identified by the Massachusetts Institute of Technology in 2013.

From a business perspective, the Big Data era was envisioned as the next frontier for innovation, competition and productivity in the McKinsey report (Manyika et al. 2011) given its potential to drive business revenues and create new opportunities. For example, Najjar and Kettinger (2013) utilized a four-stage strategy to analyze how Big Data can be monetized. Monsanto's Climate Corporation uses geospatial Big Data to analyze the weather's complex and multi-layered behavior to help farmers around the world adapt to climate change (Dasgupta 2013). Manyika et al. (2011) predicted that Big Data would improve 60% of existing businesses and foster billions of dollars of new business in the next decade. In summary, the Big Data arena ushers in great opportunities and changes in digital earth arena on how we live, think and work (Mayer-Schönberger and Cukier 2013), including personalized medicine (Alyass, Turcotte, and Meyre 2015), customized product recommendations and travel options. The past few years have witnessed this transformation from concept to reality through a host of technological innovations (e.g. Uber and Wechat).

Past research on processing Big Data focused on the distributed and stream-based processing (Zikopoulos and Eaton 2011). While cloud computing emerged a bit earlier than Big Data, it is a new computing paradigm for delivering computation as a fifth utility (after water, electricity, gas and telephony) with the features of elasticity, pooled resources, on-demand access, self-service and pay-as-you-go (Mell and Grance 2011). These features enabled cloud services to be Infrastructure as a Service, Platform as a Service and Software as a Service (Mell and Grance 2011). While redefining the possibilities of geoscience and digital earth (Yang, Xu, and Nebert 2013), cloud computing has engaged Big Data and enlightened potential solutions for various digital earth problems in geoscience and relevant domains such as social science, astronomy, business and industry. The

Table 1. Addressing the Big Data challenges with cloud computing (e.g. dust storm forecasting).

Big Data\cloud computing	Elasticity	Pooled	On-demand	Self-service	Pay-as-you-go
Volume		x			x
Velocity	x		x		
Variety	x	x		x	
Veracity				x	x
Value	x		x		x

features of cloud computing and their utilization to support characteristics of Big Data are summarized (Table 1).

Using dust storm forecasting as an example (Xie et al. 2010), these storms occur rarely in a year (~1% of time) and once initiated, they develop rapidly and normally subside in hours to days. These features make it feasible to maintain small-scale forecasting with coarse resolution. But once abnormal dust concentrations are detected, large computing resources need to be engaged quickly (i.e. minutes), to assemble Big Data from weather forecasting and ground observations at different speeds and quality (Huang et al. 2013a). Such an application has all features of the 5Vs for Big Data and can be addressed by cloud computing with relevant features (Table 1, denoted ('x')): (i) data volume processed with a large pooling of computing resource; (ii) velocity of observation and forecasting, handled by elasticity and on-demand features; (iii) variety of multi-sourced inputs addressed by elasticity, pooled resources (computing and analytics) and self-service advantages; (iv) veracity of the Big Data relieved by self-service to select the best-matched services and pay-as-you-go cost model and (v) value represented as accurate forecasting with high resolution, justifiable cost and customer satisfaction with on-demand, elasticity and pay-as-you-go features of cloud computing.

For the adoption of cloud computing for other applications, a table similar to Table 1 can be constructed. On the other hand, the increasing demand for accuracy, higher resolutions and Big Data will drive the advance of cloud computing and associated technologies. Thus, cloud computing provides a new way to do business and support innovation (Zhang and Xu 2013). The integration of cloud computing, Big Data, and economy of goods and digital services have been fostering the discussion of IT-related services, a large share of our daily purchasing consumption (Huang and Rust 2013). It is proposed that these Big Data applications with 5V features and challenges are and will be driving the explosive advancements of relevant cloud computing technologies in different directions.

To capture this explosive growth of Big Data and cloud computing in the context of digital earth, this paper presents a comprehensive survey (Figure 1) of the Big Data challenges in different domains (Section 2), Big Data technology challenges (Section 3), cloud computing and relevant technology landscape (Section 4), current status of tackling Big Data problems with cloud computing (Section 5) and a research agenda (Section 6) towards innovation (Section 7).

2. Sources of Big Data

While Big Data has the feature of 5Vs, the feature-based challenges vary in different digital earth relevant domains. This section reviews relevant domain-specific Big Data challenges in the sequence of how closely are they related to geospatial principles and the importance of spatiotemporal thinking in relevant solution developments (Table 2) as also reflected in the research agenda.

2.1. Earth sciences

The advancement of sensing and computing simulation technologies enabled collection and generation of massive data sets every second at different spatiotemporal scales for monitoring, understanding and presenting complex earth systems. For example, EO collects TB of images daily (Yang et al. 2011a) with increasing spatial (e.g. sub-meter), temporal (hourly) and spectral (hundreds of bands) resolutions (Benediktsson, Chanussot, and Moon 2013). Geospatial models also generate

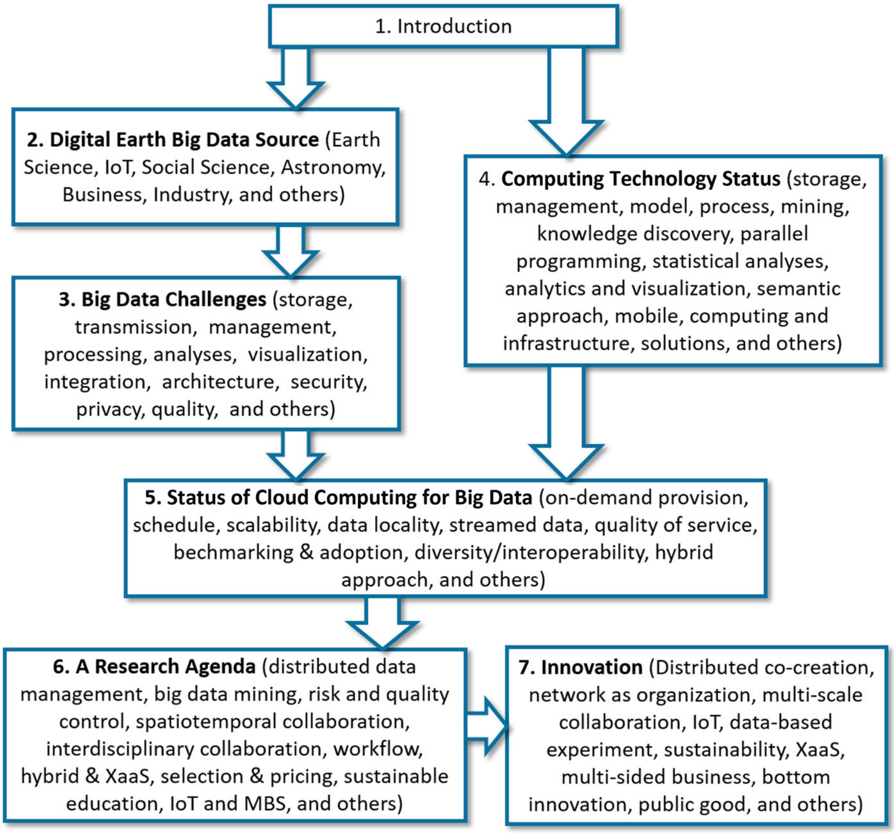
1. Introduction
2. Digital Earth Big Data Source (Earth Science, IoT, Social Science, Astronomy, Business, Industry, and others)
3. Big Data Challenges (storage, transmission, management, processing, analyses, visualization, integration, architecture, security, privacy, quality, and others)
4. Computing Technology Status (storage, management, model, process, mining, knowledge discovery, parallel programming, statistical analyses, analytics and visualization, semantic approach, mobile, computing and infrastructure, solutions, and others)
5. Status of Cloud Computing for Big Data (on-demand provision, schedule, scalability, data locality, streamed data, quality of service, bechmarking & adoption, diversity/interoperability, hybrid approach, and others)
6. A Research Agenda (distributed data management, big data mining, risk and quality control, spatiotemporal collaboration, interdisciplinary collaboration, workflow, hybrid & XaaS, selection & pricing, sustainable education, IoT and MBS, and others)
7. Innovation (Distributed co-creation, network as organization, multi-scale collaboration, IoT, data-based experiment, sustainability, XaaS, multi-sided business, bottom innovation, public good, and others)

Figure 1. Tackling Big Data challenges with cloud computing for innovation.

Table 2. The pressing features of Big Data in different domains as denoted with an 'x'.

		Earth sciences	Internet of Things	Social sciences	Astronomy	Business	Industry
Volume		x	x	x	x	x	x
Velocity		x	x	x	x	x	x
Variety	High dimensions	x		x	x		
	Spatiotemporal	x	x	x	x	x	x
	Multisource	x	x	x		x	x
Veracity		x		x	x		
Value		x	x	x	x	x	x

large spatiotemporal data *via* numerical simulations of the complex earth systems. Climate science is an exemplar representing the Big Data shift across all digital earth domains (Edwards 2010; Schnase et al. 2014) in using big spatiotemporal data to monitor and describe the complex earth climate system. For example, the IPCC AR5 alone produced 10,000 TB of climate data, and the next IPCC will engage hundreds of PB. It is critical to efficiently analyze these data for detecting global temperature anomalies, identifying geographical regions with similar or disparate climate patterns, and investigating spatiotemporal distribution of extreme weather events (Das and Parthasarathy 2009). However, efficiently mining information from PB of climate data is still challenging (Li 2015).

2.2. Internet of Things

Advanced sensors and their hosting devices (e.g. mobile phones, health monitors) are connected in a cyber-physical system to measure time and location of humans, movement of automobiles, vibration of machine, temperature, precipitation, humidity and chemical changes in the atmosphere (Lohr 2012). The Internet of Things (IoT, Camarinha-Matos, Tomic, and Graça 2013) captures this new domain and continuously generates data streams across the globe with geographical footprints from interconnected mobile devices, personal computers, sensors, RFID tags and cameras (Michael and Miller 2013; Van den Dam 2013). Big Data generated from the various sensors of IoT contains rich spatiotemporal information. The advance of IoT and Big Data technologies presents an array of applications including better product-line management, more effective and timely criminal investigation, boosting agriculture productivity (Jiang, Fang, and Huang 2009; Hori, Kawashima, and Yamazaki 2010; Xing, Pang, and Zhang 2010; Bo and Wang 2011), and accelerating the development of smart cities (Belissent 2010; Schaffers et al. 2011; Gubbi et al. 2013) with new architecture (Balakrishna 2012; Mitton et al. 2012; Theodoridis, Mylonas, and Chatzigiannakis 2013; Jin et al. 2014).

2.3. Social sciences

Social networks, such as Twitter and Facebook, generate Big Data and are transforming social sciences. As of the time of writing, Twitter users around the world are producing around 6000 tweets per second which corresponds to 500 million tweets per day and around 200 billion tweets per year (Internet Live Stats 2016). Economists, political scientists, sociologists and other social scholars use Big Data mining methods to analyze social interactions, health records, phone logs, government records and other digital traces (Boyd and Crawford 2012). While such mining methods benefit governments and social studies (Grimmer 2015), it is still challenging to quickly extract spatiotemporal patterns from big social data to, for example, help predict criminal activity (Heffner 2014), observe emerging public health threats and provide more effective intervention (Lampos and Cristianini 2010; Jalali, Olabode, and Bell 2012).

2.4. Astronomy

Astronomy is producing a spatiotemporal map of the universe (Dillon 2015) by observing the sky using advanced sky survey technologies. Mapping and surveying the universe generates vast amounts of spatiotemporal data. For example, Sloan Digital Sky Survey (SDSS) generated 116 TB data for its Data Release 12 (Alam et al. 2015). New observational instruments (e.g. Large Synoptic Survey Telescope) scheduled for operation in 2023 will generate 15 TB data nightly and deliver 200 PB data in total to address the structure and evolution of the universe (http://www.lsst.org). Besides observational data, the Large Hadron Collider is investigating how the universe originated and operates at the atomic level (Bird 2011) and produces 60 TB of experimental data per day and 15 PB data per annum (Bryant, Katz, and Lazowska 2008). Subsequent to collection, the foremost challenge in this new astronomical arena is managing the Big Data, making sense of the information efficiently, and finding interesting celestial objects and processing in an effective manner (Jagadish et al. 2014). The astronomy big data not only records information on how universe evolves, but also can be used to understanding how Earth evolves and protecting Earth from outer space impact, such as planetary defense.

2.5. Business

Business intelligence and analytics are enhanced with Big Data for decisions on strategy, managing optimization and competition (Chen, Chiang, and Storey 2012; Gopalkrishnan et al. 2012). Business actions (e.g. credit card transitions, online purchases) generate large volume, high velocity and highly unstructured (variety and veracity) data sets. These data contain rich geospatial information, such as where and when a transition occurred. To manage and process these data, the full spectrum of data processing technologies has been developed for the distributed and scalable storage environment (Färber et al. 2012; Indeck and Indeck 2012; Moniruzzaman and Hossain 2013). However, it remains a challenge to efficiently construct spatiotemporal statistical models from business data to optimize product placement, analyze customer transaction and market structure, develop personalized product recommendation systems, manage risks and support timely business decisions (Bryant, Katz, and Lazowska 2008; Hsu 2008; Duan, Street, and Xu 2011).

2.6. Industry

In the fourth industrial revolution (Industry 4.0), products and production systems leverage IoT and Big Data to build ad-hoc networks for self-control and self-optimization (O'Donovan et al. 2015). Big Data poses a host of challenges to Industry 4.0, including the following: (i) seamless integration of energy and production; (ii) centralization of data correlations from all production levels; (iii) optimization of performance of scheduling algorithms (Sequeira et al. 2014; Gui et al. 2016); (iv) storage of Big Data in a semi-structured data model to enable real-time queries and random access without time-consuming operations and data joins (Kagermann et al. 2013) and (v) realization of on-the-fly analysis to help organizations react quickly to unanticipated events and detect hidden patterns that compromise production efficiency (Sequeira et al. 2014). Cloud computing could be leveraged to tackle these challenges in Industry 4.0 for networking, data integration, data analytics (Gölzer, Cato, and Amberg 2015) and intelligence for Cyber-Physical Systems and resiliency and self-adaptation (Krogh 2008).

In addition to the reviewed six sources, Big Data challenges may also come from other relevant domains such as medical research, public health, smart cities, security management, emergency response and disaster recovery.

3. Big Data technology challenges

While following the life cycle challenges of traditional data, digital earth Big Data poses other technological challenges because of its 5V features in many different sectors of industry, government and the sciences (McAfee et al. 2012). This section reviews the technological challenges posed by Big Data.

3.1. Data storage

Storage challenges are posed by the volume, velocity and variety of Big Data. Storing Big Data on traditional physical storage is problematic as hard disk drives (HDDs) often fail, and traditional data protection mechanisms (e.g. RAID or redundant array of independent disks) are not efficient with PB-scale storage (Robinson 2012). In addition, the velocity of Big Data requires the storage systems to be able to scale up quickly which is difficult to achieve with traditional storage systems. Cloud storage services (e.g. Amazon S3, Elastic Block Store or EBS) offer virtually unlimited storage with high fault tolerance which provides potential solutions to address Big Data storage challenges. However, transferring to and hosting Big Data on the cloud is expensive given the size of data volume (Yang, Xu, and Nebert 2013). Principles and algorithms, considering the spatiotemporal patterns of data usage, need to be developed to determine the data's analytical value and its preservation datasets by balancing the cost of storage and data transmission with the fast accumulation of Big Data (Padgavankar and Gupta 2014).

3.2. Data transmission

Data transmission proceeds in different stages of data life cycle as follows: (i) data collection from sensors to storage; (ii) data integration from multiple data centers; (iii) data management for transferring the integrated data to processing platforms (e.g. cloud platforms) and (iv) data analysis for moving data from storage to analyzing host (e.g. high performance computing (HPC) clusters).

Transferring large volumes of data poses obvious challenges in each of these stages. Therefore, smart preprocessing techniques and data compression algorithms are needed to effectively reduce the data size before transferring the data (Yang, Long, and Jiang 2013). For example, Li et al. (2015a) proposed an efficient network transmission model with a set of data compression techniques for transmitting geospatial data in a cyberinfrastructure environment. In addition, when transferring Big Data to cloud platforms from local data centers, how to develop efficient algorithms to automatically recommend the appropriate cloud service (location) based on the spatiotemporal principles to maximize the data transfer speed while minimizing cost is also challenging.

3.3. Data management

It is difficult for computers to efficiently manage, analyze and visualize big, unstructured and heterogeneous data. The variety and veracity of Big Data are redefining the data management paradigm, demanding new technologies (e.g. Hadoop, NoSQL) to clean, store, and organize unstructured data (Kim, Trimi, and Chung 2014). While metadata are essential for the integrity of data provenances (Singh et al. 2003; Yee et al. 2003), the challenge remains to automatically generate metadata to describe Big Data and relevant processes (Gantz and Reinsel 2012; Oguntimilehin and Ademola 2014). Generating metadata for geospatial data is even challenging due to the data's intrinsic characteristics of high-dimensionality (3D space and 1D time) and complexity (e.g. space–time correlation and dependency). Besides metadata generation, Big Data also poses challenges to database management systems (DBMSs) because traditional RBDMSs lack scalability for managing and storing unstructured Big Data (Pokorny 2013; Chen et al. 2014a). While non-relational (NoSQL) databases such as MongoDB and HBase are designed for Big Data (Han et al. 2011; Padhy, Patra, and Satapathy 2011), how to tailor these NoSQL databases to handle geospatial Big Data by developing efficient spatiotemporal indexing and querying algorithms is still a challenging issue (Whitman et al. 2014; Li et al. 2016a).

3.4. Data processing

Processing large volumes of data requires dedicated computing resources and this is partially handled by the increasing speed of CPU, network and storage (Bertino et al. 2011). However the computing resources required for processing Big Data far exceed the processing power offered by traditional commuting paradigms (Ammn and Irfanuddin 2013). Cloud computing offers virtually unlimited and on-demand processing power as a partial solution. However, shifting to the cloud ushers in a number of new issues. First is the limitation of cloud computing's network bandwidth which impacts the computation efficiency over large data volumes (Bryant, Katz, and Lazowska 2008). Second is data locality for Big Data processing (Yang, Xu, and Nebert 2013). While 'moving computation to data' is a design principle followed by many Big Data processing platforms, such as Hadoop (Ding et al. 2013), the virtualization and pooled nature of cloud computing makes it a challenging task to track and ensure data locality (Yang, Long, and Jiang 2013), and to support data processing involving intensive data exchange and communication (Huang et al. 2013b).

In addition, the veracity of Big Data requires preprocessing before conducting data analysis and mining (e.g. cluster analysis, classification, machine learning) for better quality (LaValle et al. 2013; Mayer-Schönberger and Cukier 2013). Large, high-dimensional spatiotemporal data cannot be managed by existing data reduction algorithms within a tolerable time frame and acceptable quality (Aghabozorgi, Seyed Shirkhorshidi, and Ying Wah 2015; García, Luengo, and Herrera 2015). For example, traditional algorithms are not able to preprocess the massive volumes of continuously incoming intelligence and surveillance sensor data in real time. Highly efficient and scalable data reduction algorithms are required for removing the potentially irrelevant, redundant, noisy and misleading data, and this is one of the most important tasks in Big Data research (Zhai, Ong, and Tsang 2014).

3.5. Data analysis

Data analysis is an important phase in the value chain of Big Data for information extraction and predictions (Fan and Liu 2013; Chen et al. 2014b). However, analyzing Big Data challenges the complexity and scalability of the underlying algorithms (Khan et al. 2014). Big Data analysis requires sophisticated scalable and interoperable algorithms (Jagadish et al. 2014) and is addressed by welding analysis programs to parallel processing platforms (e.g. Hadoop) to harness the power of distributed processing. However, this 'divide and conquer' strategy does not work with deep and multi-scale iterations (Chen and Zhang 2014) that are required for most geospatial data analysis/mining algorithms. Furthermore, most existing analytical algorithms require structured homogeneous data and have difficulties in processing the heterogeneity of Big Data (Bertino et al. 2011). This gap requires either new algorithms that cope with heterogeneous data or new tools for preprocessing data to make them structured to fit existing algorithms. In geospatial domain, optimizing existing spatial analysis algorithms by integrating spatiotemporal principles (Yang et al. 2011b) to accelerate geospatial knowledge discovery is challenging and has become a high priority research field of 'spatiotemporal thinking, computing and applications' (Cao, Yang, and Wong 2009; Yang 2011; Yang et al. 2014; Li et al. 2016a).

3.6. Data visualization

Big Data visualization uncovers hidden patterns and discovers unknown correlations to improve decision-making (Nasser and Tariq 2015). Since Big Data is often heterogeneous in type, structure and semantics, visualization is critical to make sense of Big Data (Chen et al. 2014b; Padgavankar and Gupta 2014). But it is difficult to provide real-time visualization and human interaction for visually exploring and analyzing Big Data (Sun et al. 2012; Jagadish et al. 2014; Nasser and Tariq 2015). The SAS (2012) summarized five key functionalities for Big Data visualization as follows: (i) highly interactive graphics incorporating data visualization best practices; (ii) integrated, intuitive and approachable visual analytics; (iii) web-based interactive interfaces to preview, filter or sample data prior to visualizations; (iv) in-memory processing and (v) easily distributed answers and insight *via* mobile devices and web portals. Designing and developing these functionalities is challenging because of the many features of Big Data including the fusion of multiple data sources and the high-dimensionality and high spatial resolution of geospatial data (Fox and Hendler 2011; Reda et al. 2013).

3.7. Data integration

Data integration is critical for achieving the 5th V (value) of Big Data through integrative data analysis and cross-domain collaborations (Chen et al. 2013; Christen 2014). Dong and Divesh (2015) summarized the data integration challenges of schema mapping, record linkage and data fusion. Metadata is essential for tracking these mappings to make the integrated data sources 'robotically' resolvable and to facilitate large-scale analyses (Agrawal et al. 2011). However efficiently and automatically creating metadata from Big Data is still a challenging task (Gantz and Reinsel 2011). In geospatial domain, geo-data integration has sparked new opportunities driven by ever increasingly collaborative research environment. One example is the EarthCube program initiated by the U.S. NSF's Geosciences Directorate to provide unprecedented integration and analysis of geospatial data from a variety of geoscience domains (EarthCube 2014).

3.8. Data architecture

Big Data is gradually transforming the way scientific research is conducted as evidenced by the increasingly data-driven and the open science approach (Jagadish et al. 2014). Such transformations pose challenges to system architecture. For example, seamlessly integrating different tools and geospatial services remain a high priority (Li et al. 2011; Wu et al. 2011). Additional priority issues include integrating these tools into reusable workflows (Li et al. 2015b), incorporating data with the tools to promote functionality (Li et al. 2014) and sharing data and analyses among communities. An ideal architecture would seamlessly synthesize and share data, computing resources, network, tools, models and, most importantly, people (Wright and Wang 2011). Geospatial cyberinfrastructure is actively investigated in the geospatial sciences (Yang et al. 2010). EarthCube, though still in an early development stage, is a good example of such cyberinfrastructure in geospatial domain. Building a similar cyberinfrastructure for other science domains is equally important and challenging.

3.9. Data security

The increasing dependence on computers and Internet over the past decades makes businesses and individuals vulnerable to data breach and abuse (Denning and Denning 1979; Abraham and Paprzycki 2004; Redlich and Nemzow 2006). Big Data poses new security challenges for traditional data encryption standards, methodologies and algorithms (Smid and Branstad 1988; Coppersmith 1994; Nadeem and Javed 2005). Previous studies of data encryption focused on small-to-medium-size data, which does not work well for Big Data due to issues of the performance and scalability (Chen et al. 2014b). In addition, data security policies and schemes to work with the structured data stored in conventional DBMS are not effective in handling highly unstructured, heterogeneous data (Villars, Olofson, and Eastwood 2011). Thus, effective policies for data access control and safety management need to be investigated in Big Data and these need to incorporate new data management systems and storage structures (Cavoukian and Jonas 2012; Chen et al. 2014a). In the cloud era, since data owners have limited control on virtualized storage, ensuring data confidentiality, integrity and availability becomes a fundamental concern (Kaufman 2009; Wang et al. 2009; Feng et al. 2011; Chen and Zhao 2012).

3.10. Data privacy challenges

The unprecedented networking among smart devices and computing platforms contributes to Big Data but poses privacy concerns where an individual's location, behavior and transactions are digitally recorded (Cukier 2010; Tene 2011; Michael and Miller 2013; Cheatham 2015). For example, social media and individual medical records contain personal health information raising privacy concerns (Terry 2012; Kaisler et al. 2013; Michael and Miller 2013; Padgavankar and Gupta 2014). Another example is that companies are using Big Data to monitor workforce performance by tracking the employees' movement and productivity (Michael and Miller 2013). These privacy issues expose a gap between the convention policies/regulations and Big Data and call for new policies to address comprehensively privacy concerns (Khan et al. 2014; Eisenstein 2015).

3.11. Data quality

Data quality includes four aspects: accuracy, completeness, redundancy and consistency (Chen et al. 2014b). The intrinsic nature of complexity and heterogeneity of Big Data makes data accuracy and completeness difficult to identify and track, thus increasing the risk of 'false discoveries' (Lohr 2012). For example, social media data are highly skewed in space, time and demographics, and location accuracy varies from meters to hundreds of kilometers. In addition, data redundancy control and filtering should be conducted at the point of data collection in real-time (e.g. with sensor networks, Cuzzocrea, Fortino, and Rana 2013; Chen et al. 2014a). Finally, ensuring data consistency and integrity is challenging with Big Data especially when the data change frequently and are shared with multiple collaborators (Khan et al. 2014).

4. Cloud computing and other relevant technology landscape

This section reviews the technology challenges posed by Big Data from 12 different aspects. While some of these challenges (such as analysis, visualization and quality) exist before Big Data era, the 5Vs of Big Data bring the challenges to a new level as discussed above. Big Data poses unique challenges from several aspects including analysis, visualization, integration and architecture, due to the inherent high-dimensionality of geospatial data and the complex spatiotemporal relationships.

To address Big Data challenges (Sections 2 and 3), a variety of methodologies, techniques and tools (Table 3) are identified to facilitate the transformation of data into value. Computing infrastructure, especially cloud computing, plays a significant role in information and knowledge extraction. Efficient handling of Big Data often requires specific technologies, such as massive parallel processing, distributed databases, data-mining grids, scalable storage systems, and advanced computing architectures, platforms, infrastructures and frameworks (Cheng, Yang, and Rong 2012; Zhang, Li, and Chen 2012). This section introduces these methodologies and technologies that underpin Big Data handling (Table 3).

4.1. Data storage, management and model

4.1.1. Distributed file/storage system

To meet the storage challenge, an increasing number of distributed file systems (DFSs) are adapted with storage of small files, load balancing, copy consistency and de-duplication (Zhang and Xu 2013) in a network-shared files and storage fashion (Yeager 2003). The Hadoop Distributed File System (HDFS; Shvachko et al. 2010) is such a system running on multiple hosts, and many IT companies, including Yahoo, Intel and IBM, have adopted HDFS as the Big Data storage technology. Many popular cloud storage services powered by different DFSs, including Dropbox, iCloud, Google Drive, SkyDrive and SugarSync, are widely used by the public to store Big Data and overcome limited data storage on a single computer (Gu et al. 2014).

4.1.2. NoSQL database system

While contemporary data management solutions offer limited integration capabilities for the variety and veracity of Big Data, recent advances in cloud computing and NoSQL open the door for new solutions (Grolinger et al. 2013). The NoSQL databases match requirements of Big Data with high scalability, availability and fault tolerance (Chen et al. 2014a). Many studies have investigated emerging Big Data technologies (e.g. MapReduce frameworks, NoSQL databases) (Burtica et al. 2012). A Knowledge as a Service (KaaS) framework is proposed for disaster cloud data management, in which data are stored in a cloud environment using a combination of relational and NoSQL databases (Grolinger et al. 2013). Currently, NoSQL systems for interactive data serving environments and large-scale analytical systems based on MapReduce (e.g. Hadoop) are widely adopted for Big Data management and analytics (Witayangkurn, Horanont, and Shibasaki 2013). For example, the open-source Hive project integrates declarative query constructs into MapReduce-like software to allow greater data independence, code reusability and automatic query optimization.[1] HadoopDB (Abouzeid et al. 2009) incorporates Hadoop and open-source DBMS software for data analysis, achieving the performance and efficiency of parallel databases yet still achieving the scalability, fault tolerance, and flexibility of MapReduce-based systems.

4.1.3. Search, query, indexing and data model design

Performance is critical in Big Data era, and accurately and quickly locating data requires a new generation of search engines and query systems (Miyano and Uehara 2012; Aji et al. 2013). Zhong et al. (2012) proposed an approach to provide efficient spatial query processing over big spatial data and numerous concurrent user queries. This approach first organizes spatial data in terms of geographic proximity to achieve high Input/Output (I/O) throughput, then designs a two-tier distributed spatial

Table 3. Addressing the Big Data challenges with emerging methodologies, technologies and solutions.

Technology/Big Data challenge	Storage	Transfer	Management	Preprocessing/ processing	Analysis	Visualization	Integration	Architecture	Security/ privacy	Quality	Cost/energy efficiency
Distributed file/storage system	x		x								
NoSQL database systems	x		x								
Search, query, indexing and data model design	x	x	x	x	x						
MapReduce (Hadoop) system				x	x	x	x	x			
Parallel programming languages				x	x	x	x				
Statistical analyses, machine learning and data mining					x						
Big Data analytics and visualization					x	x					
Semantics			x		x		x		x	x	
Mobile data and computing		x		x	x						x
Internet of Things		x		x							
Near field communication		x									
Computing infrastructure			x	x	x	x	x				x
Cloud computing	x	x	x	x	x	x	x	x		x	x
Management and processing architecture	x		x	x	x	x	x	x			
Remote collaboration	x	x	x	x	x	x	x	x	x		
Cloud monitoring and tracking											x
Anything as a Service	x	x	x	x	x	x	x	x			
Resource auto-provision, scaling and scheduling	x		x	x	x	x	x	x			x
Spatiotemporal optimization	x	x	x	x	x	x	x	x			x

index for pruning the search space, and finally uses an 'indexing + MapReduce' data processing architecture for efficient spatial query. Aji et al. (2013) developed Hadoop-GIS, a scalable and high performance spatial data warehousing system on Hadoop, to support large-scale spatial queries. SpatialHadoop (Eldawy and Mokbel 2013) is an efficient MapReduce framework for spatial data queries and operations by employing a simple spatial high-level language, a two-level spatial index structure and basic spatial components built inside the MapReduce layer.

In fact, comprehensive and spatiotemporal indices are needed and developed to query and retrieve Big Data (Zhao et al. 2012). An example is the global B-tree or spatiotemporal indices for HDFS. Xia et al. (2014) proposed a new indexing structure, the Access Possibility R-tree (APR-tree), to build an R-tree-based index using spatiotemporal query patterns. In a later study, Xia et al. (2015b) further proposed a new indexing mechanism with spatiotemporal patterns integrated to support Big EO global access by defining heterogeneous user queries into different categories based on spatiotemporal patterns of queries, and using different indices for each category.

The spatiotemporal Big Data poses grand challenges in terms of the representation and computation of time geographic entities and relations. Correspondingly, Chen et al. (2015) proposed spatiotemporal data model using a compressed linear reference (CLR) technique to transform time geographic entities of a road network in three-dimensional (3D) (x, y, t) space to two-dimensional (2D) CLR space. As a result, network time geographic entities can be stored and managed in a spatial database and efficient spatial operations and index structures can be employed to perform spatiotemporal operations and queries for network time geographic entities in CLR space.

While progress has been made in designing search engines, parallelizing queries, optimizing indices, and using new data models for better representing and organizing heterogeneous data in various formats, more studies are required to speed up the Big Data representation, access and retrieval under various file systems and database environments.

4.2. Data processing, mining and knowledge discovery

4.2.1. MapReduce (Hadoop) system

MapReduce is a parallel programming model for Big Data with high scalability and fault tolerance. The elegant design of MapReduce has prompted the implementation of MapReduce in different computing architectures, including multi-core clusters, clouds, Cubieboards and GPUs (Jiang et al. 2015). MapReduce also has become a primary choice for cloud providers to deliver data analytical services (Zhao et al. 2014). Many traditional algorithms and data processing in a single machine environment are transferred to the MapReduce platform (Kim 2014; Cosulschi, Cuzzocrea, and De Virgilio 2013). For example, Kim (2014) analyzed an algorithm to generate wavelet synopses on the distributed MapReduce framework.

A specially designed MapReduce algorithm with grid index is proposed for polygon overlay processing, one of the important operations in GIS spatial analysis, and associated intersection computation time can be largely reduced by using the grid index (Wang et al. 2015). Cary (2011) also showed how indexing spatial data can be scaled in MapReduce for data-intensive problems. To support GIS applications efficiently, Chen, Wang, and Shang (2008) proposed a high performance workflow system MRGIS, a parallel and distributed computing platform based on MapReduce clusters. Nguyen and Halem (2011) proposed a workflow system based on MapReduce framework for scientific data-intensive workflows, and a climate satellite data-intensive processing and analysis application is developed to evaluate the workflow system.

However, many Big Data applications with MapReduce require a rapid response time, and improving the performance of MapReduce jobs is a priority for both academia and industry (Gu et al. 2014) for different objectives, including job scheduling, resource allocation, memory issues and I/O bottlenecks (Gu et al. 2014; Jiang et al. 2015).

4.2.2. Parallel programming languages

While parallel computing (e.g. MapReduce) is widely used and effective in Big Data, an urgent priority is effective programming models and languages (Hellerstein 2010; Dobre and Xhafa 2014). Alvaro et al. (2010) demonstrated that declarative languages substantially simplify distributed systems programming. Datalog-inspired languages have been explored with a focus on inherently parallel tasks in networking and distributed systems and have proven to be a much simpler family of languages for programming parallel and distributed software (Hellerstein 2010). Shook et al. (2016) developed a parallel cartographic modeling language (PCML) implemented in Python to parallelize spatial data processing by extending the widely adopted cartographic modeling framework. However, most of geoscientists and practitioners are lacking of parallel programming skills which requires considering what, when and how to parallelize an application task. Therefore, an associated parallel programming language could be created to automatically produce parallelization code with much simple and less programming work, for example, the dragging and drawing different modules or clicking a set of buttons.

4.2.3. Statistical analyses, machine learning and data mining

Standard statistical and data mining tools for traditional data sets are not designed for supporting statistical and machine learning analytics for Big Data (Triguero et al. 2015) because many traditional tools (e.g. R) only run on a single computer. Many scholars have investigated parallel and scalable computing to support the most commonly used algorithms. For example, Triguero et al. (2015) proposed a distributed partitioning methodology for nearest neighbor classification, and Tablan et al. (2013) presented a unique, cloud-based platform for large-scale, natural language processing (NLP) focusing on data-intensive NLP experiments by harnessing the Amazon cloud. Zhang et al. (2015b) proposed three distributed Fuzzy c-means (FCM) algorithms for MapReduce. In Apache Mahout, an open-source machine-learning package for Hadoop, many classic algorithms for data mining (e.g. naïve Bayes, Latent Dirichlet Allocation, LDA, logistic regression) are implemented as MapReduce jobs.

Novel approaches are also developed for big spatial data mining. Du, Zhu, and Qi (2016), for example, introduced an interactive visual approach to detect spatial clusters from emerging spatial Big Data (e.g. geo-referenced social media data) using dynamic density volume visualization in a 3D space (two spatial dimensions and a third temporal or thematic dimension of interest). Liu et al. (2016) proposed an unsupervised land use classification method with a new type of place signature based on aggregated temporal activity variations and detailed spatial interactions among places derived from the emerging Big Data, such as mobile phone records and taxi trajectories. Lary et al. (2014) presented a holistic system called Holistics 3.0 that combines multiple big datasets and massively multivariate computational techniques, such as machine learning, coupled with geospatial techniques, for public health.

However, these systems and tools still lack the features of advanced machine learning, statistical analysis and data mining (Wang, Handurukande, and Nassar 2012). To extract meaningful information from the massive data, much effort should be devoted to develop comprehensive libraries and tools that are easy to use, capable of mine massive, multidimensional (especially temporal dimension) data.

4.2.4. Big Data analytics and visualization

Big Data analytics is an emerging research topic with the availability of massive storage and computing capabilities offered by advanced and scalable computing infrastructures. Baumann et al. (2016) introduced the EarthServer, a Big Earth Data Analytics engine, for coverage-type datasets based on high performance array database technology, and interoperable standards for service interaction (e.g. OGC WCS and WCPS). The EarthServer provided a comprehensive solution from query languages to mobile access and visualization of Big Earth Data. Using sensor web event detection and geoprocessing workflow as a case study, Yue et al. (2015) presented a spatial data infrastructure (SDI) approach to support the analytics of scientific and social data.

Visual analytics also emerges as a research topic to support scientific explorations of large-scale multidimensional data. Sagl, Loidl, and Beinat (2012), for example, presented a visual analytics approach for deriving spatiotemporal patterns of collective human mobility from a vast mobile network traffic data set by relying entirely on visualization and mapping techniques.

Several tools and software have also been developed to support visual analytics, and to deliver deeper understanding. For example, the EGAN software was implemented to transform high-throughput, analytical results into a hypergraph visualizations (Paquette and Tokuyasu 2011). Gephi (http://gephi.github.io/) is an interactive visualization and exploration tool used to explore and manipulate networks and create dynamic and hierarchical graphs. Zhang et al. (2015c) presented an interactive spatial data analysis and visualization system, TerraFly GeoCloud, to help end users visualize and analyze spatial data and share the analysis results through URLs (Zhang et al. 2015b). While progress has been made to leverage cloud infrastructure and data warehouse for Big Data visualization (Edlich, Singh, and Pfennigstorf 2013), it remains a challenge to support efficient and effective exploration of Big Data, especially for dynamic and hierarchical graphs, and social media data (Huang and Xu 2014).

4.2.5. Semantics and ontology-driven approaches

Semantic and ontologies make computer and web smarter to understand, manipulate and analyze a variety of data. Semantic has emerged as a common, affordable data model that facilitates interoperability, integration and monitoring of knowledge-based systems (Kourtesis, Alvarez-Rodríguez, and Paraskakis 2014). In recent years there has been an explosion of interest using semantics for the traditional data analysis since it helps get the real information from data. Especially for cross-domain data, semantic provides advantages to link data and interchange information. Semantic understanding of Big Data is one of the ultimate goals for scientists and is used in Big Data for taxonomies, relationships and knowledge representation (Jiang et al. 2016). Currently, research is being conducted to discover and utilize semantic and knowledge database to manage and mine Big Data. For example, Liu et al. (2014) improved the recall and precision of geospatial metadata discovery based on Semantic Web for Earth and Environmental Terminology (SWEET) ontology, concept and relationships. Similar technologies could be used not only for geospatial metadata discover but also in other domains. Dantan, Pollet, and Taibi (2013) developed a semantic-enabled approach to establish a common form to design research models. Choi, Choi, and Kim (2014) proposed an ontology-based, access control model (Onto-ACM) to address the differences in the permitted access control among service providers and users. Semantic and ontological challenges remain for Big Data as the availability of content, ontology, development, evolution, scalability, multi-linguality and visualization are significantly increased (Benjamins et al. 2002).

4.3. Mobile data collection, computing and Near Field Communication

Mobile phones are playing a significant role in our lives, and mobile devices allow applications to collect and utilize Big Data and intensive computing resources (Soyata et al. 2012). In fact, large volumes of data on massive scales have been generated from GPS-equipped mobile devices, and associated methodologies and approaches are widely developed to incorporate such data for various applications, for example, human mobility studies, route recommendation, urban planning and traffic management. Huang and Wong (2016), for example, developed an approach to integrate geo-tagged tweets posted from smartphones and tradition survey data from the American Community Survey to explore the activity patterns Twitter users with different socioeconomic status. Shelton, Poorthuis, and Zook (2015) also provided a conceptual and methodological framework for the analysis of geo-tagged tweets to explore the spatial imaginaries and processes of segregation and mobility.

Sainio, Westerholm, and Oksanen (2015) introduced a fast map server to generate and visualize heat maps of popular routes online based on client preferences from massive track data collected by the Sports Tracker mobile application. Toole et al. (2015) presented a flexible, modular and computationally efficient software system to transform and integrate raw, massive data of call detail records (CDRs) from mobile phones into estimates of travel demand.

Meanwhile, mobile devices are increasingly used to perform extensive computations and store Big Data. Cloud-based Mobile Augmentation (CMA), for example, recently emerged as the state-of-the-art mobile computing model to increase, enhance and optimize computing and storage capabilities of mobile devices. For example, Abolfazli et al. (2014) surveyed the mobile augmentation domain and Soyata et al. (2012) proposed the MObile Cloud-based Hybrid Architecture (MOCHA) for mobile-cloud computing applications including object recognition in the battlefield. Han, Liang, and Zhang (2015) discussed integrating mobile sensing and cloud computing to form the singular idea of mobile cloud sensing. Edlich, Singh, and Pfennigstorf (2013) developed an open-data platform to provide Data-as-a-Service (DaaS) in the cloud for mobile devices. However, mobile devices are still limited by either capacities (e.g. CPU, memory, battery) or network resources. Solutions have been investigated to connect mobile devices to other resources with more powerful computing or network capabilities, including computers and laptops to strengthen their ability to perform computing tasks (Hung, Tuan-Anh, and Huh 2013). Near Field Communication (NFC, Coskun, Ozdenizci, and Ok 2013) allows data exchange among terminals at close distances, data transmission at low power and with a bi-directional communication and more secure transmission (Joo, Hong, and Kim 2012). Most manufacturers (e.g. Apple, Samsung, Nokia) enhance the number of mobile phones with NFC, facilitating visualization of content collected via smart media and multimedia using NFC technology (Joo, Hong, and Kim 2012).

The massive growth of mobile devices significantly changes our computer and Internet usage, along with the dramatic development of mobile services, or mobile computing. Therefore, latest networking (e.g. NFC) and computing paradigms (e.g. cloud computing) should be leveraged to enable mobile devices to run applications of Big Data and intensive computing (Soyata et al. 2012).

4.4. Big Data computing and processing infrastructure

The features of Big Data drive research to discover comprehensive solutions for both computing and data processing, such as designing advanced architectures, and developing data partition and parallelization strategies to better leverage HPC. Liu (2013) surveyed computing infrastructure for Big Data processing from aspects of architectural, storage and networking challenges and discussed emerging computing infrastructure and technologies. Nativi et al. (2015) discussed the general strategies, focusing on the cloud-based discovery and access solutions, to address Big Data challenges of Global Earth Observation System of Systems (GEOSS). Zhao et al. (2015) reviewed the current state of parallel GIS with respect to parallel GIS architectures, parallel processing strategies, relevant topics, and identified key problems and future potential research directions of parallel GIS. In fact, the widespread adoption of heterogeneous computing architecture (e.g. CPU, GPU) and infrastructure (e.g. local clusters, cloud computing) has substantially improved the computing power and raised optimization issues related to the processing of task streams across different architectures and infrastructures (Campa et al. 2014). The following sections discuss various advancements in computing and processing for Big Data.

4.4.1. Computing infrastructure

Parallel computing is one of the most widely used computing solutions to address the computational challenges of Big Data (Lin et al. 2013) through HPC cluster, supercomputer, or computing resources geographically distributed in different data centers (Huang and Yang 2011). To speed up the computation, this computing paradigm partitions a serial computation task into subtasks and uses multiple resources to run subtasks in parallel by leveraging different levels of parallelization from multi-cores, many cores, server rack, racks in a data center and globally shared infrastructure over the Internet (Liu 2013). For example, unmanned aerial vehicles (UAVs) images are processed efficiently to support automatic 3D building damage detection after an earthquake using a parallel processing approach (Hong et al. 2015). Pijanowski et al. (2014) redesigned a Land Transformation Model to make it capable of running at large scales and fine resolution (30 m) using a new architecture that employs an HPC cluster.

Distributed computing addresses geographically dispersed Big Data challenges (Cheng, Yang, and Rong 2012) where applications are divided into many subtasks. A variety of distributed computing mechanisms (e.g. HPC, Grid) distribute user requests and achieve optimal resource scheduling by implementing distributed collaborative mechanisms (Huang and Yang 2011; Gui et al. 2016). Volunteer computing has become a popular solution for Big Data, and large-scale computing and support volunteers around the world contribute their computing resources (e.g. personal computers) and storage to help scientists run applications (Anderson and Fedak 2006). Compared to HPC clusters or grid computer pool, computing resources from volunteers are free but are much less reliable for tasks with specific deadlines. For no time-constraint tasks (e.g. QMachine, weather@home, Massey et al. 2014), computing resources from volunteers is a viable and effective methodology.

Cloud computing enhances the sharing of information, applying modern analytic tools and managing controlled access and security (Radke and Tseng 2015). It provides a flexible stack of massive computing, storage and software services in a scalable manner and at low cost. As a result, more scientific applications traditionally handled by HPC or grid facilities are deployed on the cloud. Examples of leveraging cloud computing include supporting geodatabase management (Cary et al. 2010), retrieving and indexing spatial data (Wang et al. 2009), running climate models (Evangelinos and Hill 2008), supporting dust storm forecasting (Huang et al. 2013a), optimizing spatial web portals (Xia et al. 2015b) and even delivering model as a service (MaaS) for geosciences (Li et al. 2014). At the same time, various strategies and experiments are developed to better leverage cloud resources. For example, Campa et al. (2014) investigated the scaling of pattern-based, parallel applications from local CPU/GPU-clusters to a public cloud CPU/GPU infrastructure and demonstrated that CPU-only and mixed CPU/GPU computations are offloaded to remote cloud resources with predictable performance. Gui et al. (2014) recommend a mechanism for selecting the best public cloud services for different applications.

However, only limited geoscience applications have been developed to leverage the parallel computing and emerged cloud platform. Therefore, much effort should be devoted to identify applications of massive impact, of fundamental importance, and requiring the latest parallel programming and computing paradigm (Yang et al. 2011a).

4.4.2. Management and processing architecture

Big Data requires high performance data processing tools for scientists to extract knowledge from the unprecedented volume of data (Bellettini et al. 2013).To leverage cloud computing for Big Data processing, Kim, Kim, and Kim (2013) offered a framework to access and integrate distributed data in a heterogeneous environment and deployed it in a cloud environment. Somasundaram et al. (2012) proposed a data management framework for processing Big Data in the private cloud infrastructure. Similarly, Big Data processing and analyses platform are widely used for massive data storing, processing, displaying and sharing based on cloud computing (Liu 2014) or a combination of traditional clusters and cloud infrastructure (Bellettini et al. 2013), MapReduce data processing framework (Ye et al. 2012; Lin et al. 2013) and open standards and interfaces (e.g. OGC standards; Lin et al. 2013). Krämer and Senner (2015), for example, proposed a software architecture that allows for processing of large geospatial data sets in the cloud by supporting multiple data processing paradigms such as MapReduce, in-memory computing or agent-based programming. However, existing data management and processing systems still fall short in addressing many challenges in processing large-scale data, especially stream data (e.g. performance, data storage and fault tolerance) (Zhang et al. 2015a).

Therefore, an open-data platform, architecture or framework needs to be developed to leverage latest storage, computing and information technologies in managing, processing and analyzing multi-sourced, heterogeneous data in real-time.

4.4.3. Remote collaboration

To process multi-sourced, distributed Big Data, remote collaboration involving Big Data exchange and analysis among distant locations for various Computer-Aided Design and Engineering (CAD/E) applications have been explored (Belaud et al. 2014). Fang et al. (2014) discussed the need for innovation in collaboration across multi-domains of CAD/E and introduced a lightweight computing platform for simulation and collaboration in engineering, 3D visualization and Big Data management. To handle massive biological data generated by high-throughput experimental technologies, the BioExtract Server (bioextract.org) is designed as a workflow framework to share data extracts, analytic tools and workflows with collaborators (Lushbough, Gnimpieba, and Dooley 2015). Still, more studies would contribute more practical experiences to build a remote collaboration platform provides scientists an ‘easy-to-integrate’ generic tool, thus enabling worldwide collaboration and remote processing for any kind of data.

4.4.4. Cloud monitoring and tracking

Real-time monitoring of cloud resources is crucial for a variety of tasks including performance analysis, workload management, capacity planning and fault detection (Andreolini et al. 2015). Progress has been made in cloud system monitoring and tracking. For example, Yang et al. (2015a) investigated the challenges posed by industrial Big Data and complex machine working conditions and proposed a framework for implementing cloud-based, machine health prognostics. To limit computational and communication costs and guarantee high reliability in capturing relevant load changes, Andreolini et al. (2015) presented an adaptive algorithm for monitoring Big Data applications that adapts the intervals of sampling and frequency of updates to data characteristics and administrator needs. Bae et al. (2014) proposed an intrusive analyzer that detects interesting events (such as task failure) occurring in the Hadoop system.

4.4.5. Energy efficiency and cost management

Big Data coupled with cloud computing pose other challenges, notably energy efficiency and cost management. High energy consumption is a major obstacle for achieving green computing with large-scale data centers (Sun et al. 2014). Noticeable efforts are proposed to tackle this issue including GreenHDFS, GreenHadoop and Green scheduling (Kaushik and Bhandarkar 2010; Kaushik, Bhandarkar, and Nahrstedt 2010; Zhu, Shu, and Yu 2011; Goiri et al. 2012; Hartog et al. 2012). However, efficient energy management remains a challenge. While cloud computing provides virtually unlimited storage, computing and networking (Yang et al. 2011a; Ji et al. 2012; Yang, Xu, and Nebert 2013), obtaining optimal cost and high efficiency are elusive (Pumma, Achalakul, and Li 2012). Research to address this issue is emerging from different facets, including cloud cost modeling (Gui et al. 2014), resource auto- and intelligent-scaling (Röme 2010; Jam et al. 2013; Xia et al. 2015b), location-aware smart job scheduling and workflow (Mao and Humphrey 2011; Lorido-Botrán, Miguel-Alonso, and Lozano 2012; Li et al. 2015b; Gui et al. 2016) and hybrid cloud solutions (Shen et al. 2011; Bicer, Chiu, and Agrawal 2012; Xu 2012).

4.5. Big Data and cloud solutions for geospatial sciences

While cloud computing emerged as a potential solution to support computing and data-intensive applications, several barriers hinder transitioning from traditional computing to cloud computing (Li et al. 2015b). First, the learning curve for geospatial scientists is steep when it comes to understanding the services, models and cloud techniques. Second, intrinsic challenges brought by cloud infrastructure (e.g. communication overhead, optimal cloud zones) have yet to be addressed. And third, processing massive data and running models involves complex procedures and multiple tools, requiring more flexible and convenient cloud services. This section discusses the solutions used to better leverage cloud computing for addressing Big Data problems in the geospatial sciences.

4.5.1. Anything as a service (XaaS) for geospatial data and models

There are emerging features to ease the challenge of leveraging cloud computing and facilitate the use of widely available models and services. Notable examples are web service composition (Liu et al. 2015), web modeling (Geller and Turner 2007), application as a service (Lushbough, Gnimpieba, and Dooley 2015), MaaS (Li et al. 2014) and workflow as a service (WaaS; Krämer and Senner 2015). The effectiveness of these features is based on 'hiding' from the scientific community the complexity of computing, data and the model and providing an easy-to-use interface for accessing the underlying modeling and computing infrastructure. Web and model service dynamic composition is a key technology and a reliable method for creating value-added services by composing available services and applications (Liu et al. 2015). Chung et al. (2014) presented CloudDOE to encapsulate technical details behind a user-friendly graphical interface, thus liberating scientists from performing complicated operational procedures. Uncinus allows researchers easy access to cloud computing resources through web interfaces (Lushbough, Gnimpieba, and Dooley 2015). By utilizing cloud computing services, Li et al. (2014) extended the traditional Model Web concept (Geller and Turner 2007) by proposing MaaS.

Krämer and Senner (2015) proposed a software architecture containing a web-based user interface where domain experts (e.g. GIS analysts, urban planners) define high-level processing workflows using a domain-specific language (DSL) to generate a processing chain specifying the execution of workflows on a given cloud infrastructure according to user-defined constraints. The proposed architecture is WaaS. Recent advancements in cloud computing provide novel opportunities for scientific workflows. Li et al. (2015b) offered a scientific workflow framework for big geoscience data analytics by leveraging cloud computing, MapReduce and Service-Oriented Architecture (SOA). Zhao et al. (2012) presented the Swift scientific workflow management system as a service in the cloud. Their solution integrates Swift with the OpenNebula Cloud platform. Chung et al. (2014) presented CloudDOE, a platform-independent software in Java to encapsulate technical details behind a user-friendly graphical interface.

However, general cloud platforms are not designed to support and integrate geoscience algorithms, applications and models, which might be data, computation, communication or user intensive, or a combination of two or even three of these intensities (Yang et al. 2011a). Therefore, an integrated framework should be developed to enable XaaS that can synthesize computing and model resources across different organizations, and deliver these resources to the geoscience communities to explore, utilize and leverage them efficiently and easily.

4.5.2. Computing resource auto-provision, scaling and scheduling

The large-scale data requirement of collaboration, efficient data processing and the ever increasing disparate forms of user applications are not only making data management more complex but are bringing more challenges for the resource auto-provision, scaling and scheduling of the beneath computing infrastructure (Gui et al. 2016). With the development of cloud technologies and extensive deployment of cloud platforms, computing facilities are mostly created in the format of virtualized resources or virtual machines (VMs). The implementation of an efficient and scalable computing platform with such resources is a challenge and important research topic than the traditional computing environments (Wu, Chen, and Li 2015; Li et al. 2016b). Many strategies and algorithms are proposed for cloud resource optimization, scalability and job/workflow improvement, and these are addressed in Sections 5.1, 5.2 and 5.3.

4.5.3. Optimizing Big Data platforms with spatiotemporal principles

The optimization strategy for designing, developing and deploying geospatial applications using spatiotemporal principles are discussed in Yang et al. (2011a, 2011b). Spatiotemporal patterns can be observed from the following: (i) physical location of computing and storage resources; (ii) distribution of data; (iii) dynamic and massive concurrent access of users at different locations and times and (iv) study area of the applications. A key technique for making Big Data applications perform well is location, time and data awareness by leveraging spatiotemporal patterns. Kozuch et al. (2009) demonstrated that location-aware applications outperform those with no location-aware capability by factor of 3–11 in the performance. Based on context awareness of data, platform and services, Feng et al. (2014) presented a tourism service mechanism that dynamically adjusts the service mode based on the user context information and meets the demand for tourism services in conjunction with the individual user characteristics.

The 5V's Big Data challenges also call for new methods to optimize data handling and analysis. Correspondingly, Xing and Sieber (2016) proposed a land use/land cover change (LUCC) geospatial cyberinfrastructure to optimize Big Data handling at three levels: (1) employ spatial optimization with graph-based image segmentation, (2) use a spatiotemporal Atom Model to temporally optimize the image segments for LUCC and (3) spatiotemporally optimize the Big Data analysis. Other examples include: (1) Yang et al. (2014) proposed a novel technique for processing big graphical data on the cloud by compressing Big Data with its spatiotemporal features; and (2) Xia et al. (2015a) developed a spatiotemporal performance model to evaluate the quality of distributed geographic information services by integrating cloud computing and spatiotemporal principles at a global scale.

5. Current status of tackling Big Data challenges with cloud computing

While the Big Data challenges can be tackled by many advanced technologies, such as HPC, cloud computing is the most elusive and important. This section reviews the status of using cloud computing to address the Big Data challenges.

5.1. On-demand resource provision

The volume and velocity challenges of Big Data require VM creation on-demand. Autonomous detection of the velocity for provisioning VMs is critical (Baughman et al. 2015) and should consider both optimal cost and high efficiency in task execution (Pumma, Achalakul, and Li 2012). Research is being conducted to understand the applications and relevant Big Data changing patterns to form a comprehensive model to predict system behavior as the usage patterns evolve and working loads change (Castiglione et al. 2014). For example, Pumma, Achalakul, and Li (2012) proposed an automatic mechanism to allocate the optimal numbers of resources in the cloud, and Zhang et al. (2015b) proposed a task-level adaptive MapReduce framework for scaling up and down as well as effectively using computing resources. Additionally, many workload prediction and resource-allocation models and algorithms provide or improve auto-scaling and auto-provisioning capability. For example, Zhang et al. (2015a) applied two streaming, workload prediction methods (i.e. smoothing, Kalman filters) to estimate the unknown workload characteristics and achieved a 63% performance improvement. Zhang, Chen, and Yang (2015) offered a nodes-scheduling model based on Markov chain prediction for analyzing big streaming data in real time. Pop et al. (2015) posited the deadline as the limiting constraint and proposed a method to estimate the number of resources needed to schedule a set of aperiodic tasks, taking into account execution and data transfer costs.

However, more research is necessary for improving auto-scaling and auto-provisioning capability under cloud architecture to address Big Data challenges. For example, advanced VM provision strategies should be designed to handle concurrent execution of multiple applications on the same cloud infrastructure.

5.2. Scheduling

Job scheduling effectively allocates computing resources to a set of different tasks. However, scheduling is a challenge in automatic and dynamic resource provisioning for Big Data (Vasile et al. 2014; Gui et al. 2016). Zhan et al. (2015) proposed several research directions for cloud resource scheduling, including real-time, adaptive dynamic, large-scale, multi-objective, and distributed and parallel scheduling. Sfrent and Pop (2015) showed that under certain conditions one can discover the best scheduling algorithm. Hung, Aazam, and Huh (2015) presented a cost- and time-aware genetic scheduling algorithm to optimize performance. An energy-efficient, resource scheduling and optimization framework is also proposed to enhance energy consumption and performance (Sun et al. 2015). As one of the most popular frameworks for Big Data processing, Hadoop MapReduce is optimized (e.g. task partitioning, execution) to accommodate better Big Data processing (Slagter et al. 2013; Gu et al. 2014; Hsu, Slagter, and Chung 2015; Tang et al. 2015). Progress on scheduling was made in geospatial fields as well. Kim, Tsou, and Feng (2015), for example, implemented a high performance, parallel agent-based scheduling model that has the potential to be integrated in spatial decision-support systems and to help policy-makers understand the key trends in urban population, urban growth and residential segregation. However, research effort is need on developing more sophisticated scheduling algorithms that can leverage the spatial relationships of data, computing resources, application and users to optimize the task execution process and the resource utilization of the underlying computing infrastructure.

5.3. Scalability

Scalability on distributed and virtualized processors is and has been a bottleneck for leveraging cloud computing to process Big Data. Feller, Ramakrishnan, and Morin (2015) investigated cloud scalability and concluded: (i) co-existing VMs decrease the disk throughput; (ii) performance on physical clusters is significantly better than that on virtual clusters; (iii) performance degradation due to separation of the services depends on the data-to-compute ratio and (iv) application completion progress correlates with the power consumption, and power consumption is application specific. Various cloud performance benchmarks and evaluations (Nazir et al. 2012; Huang et al. 2013b) demonstrated that balancing the number and size of VMs as a function of the specific applications is critical to achieve optimal scalability for geospatial Big Data applications. Ku, Choi, and Min (2014) analyzed four cloud performance-influencing factors (i.e. number of query statements, garbage collection intervals, quantity of VM resources and virtual CPU assignment types). Accordingly, different strategies improve scalability and achieve cost-effectiveness while handle Big Data processing tasks with scalable and elastic service to disseminate data (Wang and Ma 2015). For example, Ma et al. (2015) proposed a scalable and elastic total order service for content-based publish/subscribe using the performance-aware provisioning technique to adjust the scale of servers to the churn workloads.

With the development of various clouds and each cloud has its unique advantages and disadvantages (Huang and Rust 2013; Gui et al. 2014), the hybrid cloud integrating multiple clouds (e.g. public and private clouds) would be a new trend while adopting cloud solutions to make full use of different clouds. Correspondingly, it would be also necessary to allocate computing resources and distribute the tasks in a way that improves the scalability and performance of networked computing nodes coming from a hybrid environment.

5.4. Data locality

System I/O poses a bottleneck for Big Data processing in the cloud (Kim et al. 2014) especially when data and computing are geographically dispersed. Researchers either move data to computing resource or move the computing resources to the data (Jayalath, Stephen, and Eugster 2014).

Considering the spatiotemporal collocation in the process of scheduling to allocate resources and move data significantly improves system performance and addresses I/O issues (Hammoud and Sakr 2011; Rasmussen et al. 2012; Yang et al. 2013; Kim et al. 2014; Feller, Ramakrishnan, and Morin 2015). For example, to achieve higher locality and reduce the volume of shuffled data, Kim et al. (2014) developed a burstiness-aware I/O scheduler to enhance the I/O performance up to 23%. Using data locality, grouped slave nodes, and k-means algorithm to hybrid Mapreduce clusters with low intra-communication and high intra-communication, Yang, Long, and Jiang (2013) observed a 35% performance gain. Hammoud and Sakr (2011) described Locality-Aware Reduce Task Scheduler for improving MapReduce performance. Zhang et al. (2014) presented a data-aware programming model by identifying the most frequently shared data (task- and job-level) and replicating these data to each computing node for data locality.

In many scenarios, input data are geographically distributed among data centers, and moving all data to a single center before processing is prohibitive (Jayalath, Stephen, and Eugster 2014). In such a distributed cloud system, efficiently moving the data is important to avoid network saturation in Big Data processing. For example, Nita et al. (2013) suggest a scheduling policy and two greedy scheduling algorithms to minimize individual transfer times. Sandhu and Sood (2015b) proposed a global architecture for QoS-based scheduling for Big Data applications distributed over cloud data centers at both coarse and fine grained levels.

5.5. Cloud computing for social media and other streamed data

The variety and veracity of social media and other streamed data pose new challenges to the contemporary data processing and storage frameworks and architectures. For Big Data management, many non-traditional methodologies such as NoSQL and scalable SQL are implemented (Nambiar, Chitor, and Joshi 2014). More than often, NoSQL databases, such as MongoDB and Hadoop Hive, are used to store and manage social media data as document entries instead of relational tables (Padmanabhan et al. 2013; Huang and Xu 2014). Meanwhile, to address big streaming data processing challenges, scalable distributed computing environments based on cloud computing are leveraged (Gao et al. 2014; Cao et al. 2015; Huang et al. 2015). For example, Zelenkauskaite and Simões (2014) implemented an android-based mobile application and designed a cloud architecture to perform computationally intensive operations, including searching, data mining and large-scale, data processing. Huang et al. (2015) presented a CyberGIS framework that automatically synthesizes multi-sourced data (e.g. social media, socioeconomic data) to track disaster events, produce maps, and perform spatial and statistical analysis for disaster management. The proposed framework supports spatial Big Data analytics of multiple sources. Cao et al. (2015) also presented a scalable computational framework using an Apache Hadoop cluster to process massive location-based social media data for efficient and systematic spatiotemporal data analysis. An interactive flow mapping interface supporting real-time and interactive visual exploration of movement dynamics is developed and used to demonstrate the advantages and performance of this framework. Additionally, multimedia streaming data (e.g. social media, remote sensing) are difficult to analyze and process in real time because of the rapid arriving speed and voluminous data fields. Zhang et al. (2015c) constructed a Markov chain model to predict the varying trend of big streaming data and the appropriate cloud computing nodes to process big streaming data.

5.6. Quality of Service

Quality of Service (QoS) describes the overall performance and is particularly important for Big Data applications and cloud computing in scheduling applications on the distributed cloud (Chen et al. 2013; Sandhu and Sood 2015b). If data services and cloud data centers are geographically distributed, it is essential to monitor the QoS globally for Big Data implementation and cloud computing. For example, Xia et al. (2015a) used thousands of globally distributed volunteers to monitor the OGC

Web Map Services (WMS) and Web Coverage Services (WCS). Sandhu and Sood (2015b) proposed a global architecture for QoS-based scheduling for Big Data applications distributed to different cloud data centers. Kourtesis, Alvarez-Rodríguez, and Paraskakis (2014) outlined a semantic-based framework for QoS management to leverage semantic technologies and distributed and data-streamed processing techniques. However, more efforts should be devoted to handling multiple QoS requirements from different users in the process of resource and task scheduling within a single or multiple cloud environment(s).

5.7. Cloud computing benchmark and adoption

Benchmarking helps researchers evaluate cloud computing performance for Big Data. Huang et al. (2013b) tested the CPU, memory and I/O and network performance of three open-source cloud solutions, including CloudStack, Eucalyptus and OpenNebula. Luo et al. (2012) proposed a benchmark suite CloudRank-D to benchmark and rank cloud computing systems shared for running Big Data applications. Li et al. (2014) developed a performance evaluation methodology and quantified the runtime configuration of content-driven applications. Each cloud platform has its own unique strengths and liabilities (Huang et al. 2013b). Corresponding, Gui et al. (2014) recommended a mechanism for selecting the best public cloud service at the levels of IaaS and PaaS. However, choosing the best cloud services for a specific application requires accounting for multiple factors, such as performance, price, technologies, policies, preferences and the computing requirements, and is still a challenge.

5.8. Diversity and interoperability

The growing Big Data usage in different systems introduces heterogeneity, requiring metadata-based interoperability and different standards are adopted for such interoperability. For example, GEOSS Clearinghouse uses a standard and many wrappers (ISO 19139) to describe and transform geographic data (Liu et al. 2011). Since cloud vendors are based on many different cloud services, cloud computing and Big Data must develop relevant standards to achieve interoperability among clouds.

5.9. Hybrid computing infrastructure and virtual resource bursting

The variety and veracity of Big Data and the performance, privacy and security concerns demand a comprehensive cloud solution with services from both private and public cloud offerings (Lee, Park, and Shin 2012). Banditwattanawong, Masdisornchote, and Uthayopas (2014) stored Big Data in hybrid clouds built with different public cloud providers for performance. Campa et al. (2014) proposed an integrated programming framework for both local and remote resources for the offloading of computations from structured parallel applications to heterogeneous cloud resources. Gui et al. (2014) compared different cloud information (including price, CPU and memory) to develop a portal to help consumers better select from a diverse set of computing services. Another aspect of the hybrid cloud environment is task migration from one cloud to another cloud, bursting and leveraging strengths of different cloud resources. Research is needed to optimize the solutions for general, categorical or specific applications.

6. Research agenda

Big Data techniques and relevant challenges must be underpinned by a substantive research initiative. Several investigations have identified new initiatives for Big Data and cloud computing (Bughin, Chui, and Manyika 2010; Karimi 2014; Assunção et al. 2015; Hashem et al. 2015) from different facets of management/architecture, model development, visualization, interaction and business.

To fully leverage and advance cloud computing for Big Data manipulation, research is required to progress from the volume, velocity, variety and veracity of Big Data to the value of the products, and these research initiatives are described below.

6.1. Distributed data storage and management challenges

Big Data (e.g. big sensor data) is collected in a global and geographically dispersed fashion. A fast, next generation network connecting storages is critical to manage Big Data which also requires an efficient distributed storage system (Aydin et al. 2015). The advancement of various distributed storage and processing solutions from Hive, Hadoop, to Spark have advanced our understanding, but more emphasis is needed to discover better support systems, especially for Big Data (Yang et al. 2015b; Li et al. 2016a) in the digital earth domain with spatiotemporal characteristics. There are several research questions for optimizing distributed storage management:

- How to leverage and link the distributed storages to achieve serialization with low latency in geographically dispersed storage systems (Zhang et al. 2013)?
- How to optimize different traditional (e.g. MySQL, Oracle) and emerging DBMSs (NoSQL) for distributed storage (Agrawal, Das, and El Abbadi 2011)?
- How to globally optimize the performance of cloud systems and optimum data backup and protection options in a cloud environment (Hung, Tuan-Anh, and Huh 2013; Cheng et al. 2015)?
- How to partition and allocate Big Data into a variety of storage solutions including hard drive, RAM and cache, which collectively are critical for improving system performance (Slagter, Hsu, and Chung 2015)?
- How to augment mobile storage with cloud storage using approaches to ensure optimized management and usage of distributed storage in mobile devices and the cloud (Aminzadeh, Sanaei, and Ab Hamid 2015)? and
- How to design smart storage devices that preprocess or conduct segments of data processing on the storage leveraging the co-location?

6.2. Big Data mining

Big Data (e.g. mobile sensing) require real-time data processing, information extraction and automation for decision support (Han, Liang, and Zhang 2015; Kehoe et al. 2015). Extracting values from Big Data with data mining methodologies using cloud computing require the following:

- Detect and mine outliers and hidden patterns from Big Data with high velocity and volume (Yang et al. 2015a);
- Mine geospatial and topological networks and relationships (using machine learning and intelligence) from the data of IoT (Doody and Shields 2012);
- Develop holistic research directed at the distribution of traditional data mining algorithms and tools to cloud computing nodes and centers for Big Data mining (Zhang, Chen, and Leng 2015);
- Develop a new class of scalable mining methods that embrace the storage and processing capacity of cloud platforms (Triguero et al. 2015);
- Address spatiotemporal data mining challenges in Big Data by examining how existing spatial mining techniques succeed or fail for Big Data (Vatsavai et al. 2012); and
- Provide new mining algorithms, tools and software as services in the hybrid cloud, service systems.

6.3. Risk management and quality assurance

Many data include privacy information of legal or ethical concerns (Knoppers and Chadwick 2014). It is a challenge to provide privacy protection for sensitive data using proper security and Big Data sanitization, anonymization, and generalization given the increasingly detailed, personal information nested in Big Data (Zhang et al. 2014). The risk is in the data, storage and computing/processing components of Big Data. Biased assessment from partial or limited access Big Data needs to be eliminated to reduce risk. Deploying isolation, block device encryption and two-way authentication offer solutions to ensure the confidentiality and integrity of user data (Wu, Chen, and Li 2015). Ensuring efficiency while protecting data integrity and security requires balance and must be integrated into quality of services and service level agreements (Liu et al. 2015). Privacy protection in the cloud and IoT is relevant to Big Data processing (Fu et al. 2015; Perera et al. 2015). However, fully managing and reducing risks associated with quality, privacy, security, integration and vendor for Big Data processing is a challenge.

The highest priority research initiatives to address these issues are as follows:

- Rapid identification and attack prevention using different data analytics and visualization to reduce system vulnerability (Huang, Lu, and Zhang 2015).
- Security of entering sensitive data onto untrusted cloud computing environments across geographical and jurisdiction boundaries (Liu et al. 2015);
- Fastly, efficiently and accurately locating datasets from Big Data with specific domains and a generalized conceptual framework (Megler and Maier 2012);
- Tracking and maintaining quality and trust information (Manuel 2013) given the veracity and variety characteristics of Big Data (e.g. social media data); and
- Modeling the performance of Big Data processing to ensure QoS and acceptance of service level agreements (Terry et al. 2013; Castiglione et al. 2014).

6.4. Spatiotemporal collocation

Moving data to computing has been the norm over the last several decades (Moore, Prince, and Ellisman 1998), and distributed programming models, such as MapReduce, facilitate moving computing to data (Dean and Ghemawat 2008). This collocation of data and computing is critical and the on-demand access requirement for obtaining value from Big Data, especially that in real time, adds the temporal dimension. With the features of Big Data and cloud computing, analyzing and utilizing the spatiotemporal collocation is an emerging direction and includes the following facets:

- Better management of Big Data with new spatiotemporal index for locating and extracting data in real time (Li et al. 2016a);
- Schedule Big Data applications to different cloud data centers across the globe (Sandhu and Sood 2015a);
- Develop a set of multi-criteria and a matrix to monitor and assess the dynamic workload of hybrid cloud, computing infrastructure, problem geographic locations, and data locations to collocate spatiotemporally data, computing, applications and users (Yang, Xu, and Nebert 2013); and
- Improve engineering operations and enhancing/enabling values produced for end users (Yang et al. 2011a) with innovative products and services for industry.

6.5. Interdisciplinary collaboration

Adopting cloud computing for different domains to support scientific and business needs is a long-term initiative. These cross-disciplinary efforts demand multi-scale and multi-side innovations, such as collaborative and automated coding for unstructured Big Data (Brunswicker, Bertino, and Matei 2015). Conversely, virtual co-creation and the networked community play a bigger role in accelerating innovation (Bughin, Chui, and Manyika 2010). Hiding the technical complexities of cloud computing and Big Data while exposing easy-to-use higher level services would facilitate interdisciplinary collaboration (Church, Goscinski, and Lefèvre 2015). Spatiotemporal thinking plays a critical role in this cross-domain collaborations. However, challenges exist on leveraging cloud computing to support the cross-domain Big Data applications, and the principal challenges are as follows:

- Leveraging the utility-supplied cloud services delivered over the Internet to accommodate the vast volumes of data generated by high-throughput experimental technologies;
- Adapting globally deployed cloud computing resources to integrate geospatial information from different domains of geosciences (e.g. ocean, land, atmosphere, polar) to address global challenges of climate change, and water/food shortages;
- Using cloud computing in teams with other computing resources (e.g. Hadoop, SPARC) to better support domain needs of Big Data handling for such initiatives as population and geographical data management and analyses for health and medical research (O'Driscoll et al. 2015); and
- Telling a story using Big Data and cloud computing to enhance entertainment that allow one to be genuinely connected to a virtual geographical environment by creating a far-more-engaging future entertainment connecting to the life of the audience (Schlieski and Johnson 2012).

6.6. Workflow sharing in business, science and application logic support

Workflow captures a specific application, solution logic process and relevant data for sharing among the different domains to repeat a research process or reduce investment for larger values (Van Der Aalst et al. 2003; Van Der Aalst and Van Hee 2004). Workflow is an important trend in integrating the IoT, mobile Internet, cloud computing and Big Data technologies to build a sensing environment to support global business and applications (Li et al. 2015b). Monitoring the quality of existing services globally and predicting their dynamic trend supports better workflow sharing (Liu et al. 2015; Xia et al. 2015a). The first step is enabling workflow execution on a distributed environment and leveraging features of cloud computing for Big Data (Li et al. 2015b). However, there are numerous obstacles for sharing workflow cross-domains and boundaries including the following:

- Improving cloud computing to share workflows and model the scaling capability to match the disruptive needs (e.g. dust storm forecasting) of Big Data's 5Vs (Pandey and Nepal 2013);
- Building a set of data processing support for Big Data in a Big Data as a Service (BDaaS) fashion (e.g. storage availability on the cloud across geographical regions) (Pandey and Nepal 2013);
- Developing a flexible software architecture so workflow is automatically executed with proper software services supported by relevant computing resources (Krämer and Senner 2015);
- Using web-based and service-oriented architecture for workflow system to support flexible environments for analyzing Big Data (Lushbough, Gnimpieba, and Dooley 2015); and
- Using workflow to support decision-support analytics to scale, scope and speed economies (Demirkan and Delen 2013).

6.7. Hybrid approach and anything as a service

The variety and veracity of Big Data demand support of different data, domains and locations with a comprehensive cloud computing capacity. The combination of private and public cloud to support different community and organizational needs, such as privacy and security in public health (Griebel et al. 2015), is becoming a norm for the mix of resources. The hybrid approach will also bring in more resources as a service in a cloud fashion for Big Data (e.g. manufacturing as a service as evidenced by Alibaba and Uber). A long held goal of Analytics as a Service (AaaS) or BDaaS is gradually taking shape, and innovations will be critical in this formative process (Assunção et al. 2015). For example, Zhang suggested integrating boxed solutions for sensitive or categorized Big Data processing. This innovation requires the following:

- Fostering innovation cross-cloud security, interoperability, system architecture and usage measurements as well as cross-cloud scheduling and management for optimizing task throughput and resource utilization (Pop et al. 2015);
- Managing with better reliability with comprehensive models (Tamura and Yamada 2015);
- Using dynamic hierarchies, resource-allocation to optimize Big Data processing in a cloud computing environment when engaging geographically dispersed servers and divisible applications with their workflow partitioned and executed on different cloud platforms (Ismail and Khan 2015); and
- Leveraging existing resources, including GPU and smart storage drives, to fully utilize the computing power but avoid the limited RAM and data I/O (Jiang et al. 2015).

6.8. Selection, activation and pricing

With AaaS and BDaaS, cloud computing not only provides the computing infrastructure to support Big Data but also provides a tested business model to sustain Big Data innovation. Relevant optimizational algorithms for selection and activation warrant being investigated in the context of the previously mentioned factors (Sfrent and Pop 2015) to address energy efficiency and sustainability (Sun et al. 2015), More specifically, the following initiatives are offered:

- Integrating multiple cloud services to obtain a heath ecosystem to support Big Data and other research (Pandey and Nepal 2013);
- Developing a comprehensive framework for monitoring the health of machines for both cloud and IoTs (Yang et al. 2015a);
- Developing tools to precisely measure usage of resources, including computing and data for pricing purposes (Gui et al. 2014);
- Monitoring of Big Data in the cloud to optimized selection and activation (Andreolini et al. 2015);
- Developing a comprehensive cost model and relevant automatic scheduling technologies to provide on-demand and cost-efficiency solutions for a variety of applications, including epidemic (Pumma, Achalakul, and Li 2012) and dust storm forecasting (Huang et al. 2013a);
- Developing benchmarks or criteria for modeling the quality, performance, security and integrity (Kos et al. 2015);
- Developing a scheduling protocol for resources in different aspects of adaptive dynamic, large-scale, multi-objective and distributed and parallel capabilities for optimized activation and cost saving, especially for disruptive stream data and applications (Tang et al. 2015; Zhang, Chen, and Yang 2015; Zhan et al. 2015); and
- Developing tools to incorporate spatiotemporal context of data, cloud data center, and application in the selection of cloud and Big Data services (Yang et al. 2011b).

6.9. Provide sustainable and available education

Education for the next generation of scientists, educators and decision-makers starts with the K-12 curriculum for which the paradigm is changing in part facilitated by Big Data and the easy-to-access information devices in our daily lives (Sultan 2010). As this new paradigm changes how we live, work and think (Mayer-Schönberger and Cukier 2013), learning different aspects of Big Data and understanding their impacts from analytics to intelligence will help us become smarter about Big Data, cloud computing and their applications (Chen, Chiang, and Storey 2012). Some examples of the educational role for Big Data and cloud computing are as follows:

- A full range of education to strengthen Big Data from K-12 onward to professional training, focusing on how to facilitate Big Data's application to business and including daily decision and virtual experiments using information derived from Big Data accumulated for scientific research (Dubey and Gunasekaran 2015);
- Leveraging of Big Data and cloud computing to provide more reliable and interactive e-learning experiences (Sandhu and Sood 2015a);
- Attention to social challenges to recognize and lessen differences between private and public schools and eliminate obstacles with better security, privacy protection and quality (Lim, Grönlund, and Andersson 2015); and
- Instilling the spatiotemporal thinking into K-16 for training the future work-force smarter on dealing with Big Data and optimizing cloud computing.

6.10. Internet of Things and mobile-based services

The IoT and mobile-based services support and demand innovation from the bottom of the pyramid as evidenced by China's citizen innovation movement, and the creation of public good out of Big Data (Bughin, Chui, and Manyika 2010; Zhang, Li, and Chen 2012). The mobile Internet and social networking services are emerging markets for innovations. Shekhar et al. (2012) posited that mobility and cloud computing pose technological challenges for routing and popular spatial services. Bonomi et al. (2012) proposed fog computing to address the mobility and IoT in that these possess the characteristics of low latency and location awareness, widespread geographical distribution, mobility, large number of nodes, prominent role of wireless access, strong presence of streaming and real-time applications, and heterogeneity. The advancement of IoT and mobility service is expected to bring the public within easy access to valuable results from Big Data and cloud computing. However, advancements are needed to provide the best information or knowledge enabling the public to make decisions including travel plans and health/property insurance (Abbas et al. 2015).

7. Innovation support

These research initiatives respond to the 10 aspects envisioned to produce the next generation of valuable technology-enabled businesses as identified by the McKinsey report (Bughin, Chui, and Manyika 2010; Table 4). For example, to support distributed co-creation across the computer network, the McKinsey report proposed advancements in distributed storage, interdisciplinary collaboration, workflow sharing and mobile computing as well as to collocate spatiotemporally resources and creators. Experiments are increasingly dependent on simulations facilitated by Big Data and cloud computing to address challenges in engineering design of complex systems, including health care, logistics and manufacturing (Xu et al. 2015) in the digital earth context. This requires a well-trained workforce, collaboration cross-domains, data mining analytics and spatiotemporal collocation of various data, processing and domain resources. Investing in resources to promote the public good requires support from all research directions, spatiotemporal collocation is a cardinal component to achieve all 10 innovations with methodologies, tools and solutions. For example, multi-scale collaborations require multi-spatiotemporal levels of collaboration across different domains supported by distributed storage.

The 5Vs that characterize Big Data and the five features of cloud computing increasingly play a dominant role in this innovation process for digital earth. The current innovation opportunities and research agenda of utilizing cloud computing for tackling Big Data are summarized. Many more features of this innovation are emerging with the popularization and expansion of Big Data in how we

Table 4. Research direction support for innovation.

Research/innovation	6.1 Distributed storage	6.2 Data mining	6.3 Risk and quality	6.4 Spatiotemporal collocation	6.5 Interdisciplinary collaboration	6.6 Workflow sharing	6.7 Hybrid XaaS	6.8 Pricing model	6.9 Sustainable education	6.10 IoT and mobile
1. Distributed co-creation	x			x	x	x				x
2. Network as organization				x	x	x				x
3. Multi-Scale collaborations	x			x	x					
4. IoT	x	x	x	x			x	x		x
5. Data-based experiment		x	x	x	x	x			x	
6. Sustainability	x		x	x				x		x
7. XaaS	x			x			x			
8. Multi-sided business	x	x	x	x	x		x	x		x
9. Bottom innovation	x			x	x	x		x	x	x
10. Public good	x	x	x	x	x	x	x	x	x	x

live, work, think and prosper. It is expected that the spatiotemporal thinking and research directions presented herein serve as a guide for the next decadal innovation and entrepreneurialism in relevant domains.

Note

1. http://hive.apache.org/

Acknowledgements

We thank the anonymous reviewers for their insightful comments and reviews. Dr George Taylor reviewed a previous version of this manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by NASA AIST Program [NNX15AM85G], NCCS [NNG14HH38I], Goddard [NNG16PU001], NSF I/UCRC [1338925], EarthCube [ICER-1540998], and CNS [1117300], as well as Microsoft, Amazon, Northrop Grumman, Harris, and United Nations.

ORCiD

Chaowei Yang http://orcid.org/0000-0001-7768-4066

References

Abbas, A., K. Bilal, L. Zhang, and S. U. Khan. 2015. “A Cloud Based Health Insurance Plan Recommendation System: A User Centered Approach.” *Future Generation Computer Systems* 43–44: 99–109.

Abolfazli, S., Z. Sanaei, E. Ahmed, A. Gani, and R. Buyya. 2014. “Cloud-Based Augmentation for Mobile Devices: Motivation, Taxonomies, and Open Challenges.” *IEEE Communications Surveys Tutorials* 16 (1): 337–368.

Abouzeid, A., K. Bajda-Pawlikowski, D. Abadi, A. Silberschatz, and A. Rasin. 2009. “HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads.” *Proceedings of the VLDB Endowment* 2 (1): 922–933.

Abraham, A., and M. Paprzycki. 2004. “Significance of Steganography on Data Security.” In *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC’04)*. Vol. 2, 347–351. IEEE.

Aghabozorgi, S., A. Seyed Shirkhorshidi, and T. Ying Wah. 2015. “Time-series Clustering – A Decade Review.” *Information Systems* 53 (C): 16–38.

Agrawal, D., P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, et al. 2011. *Challenges and Opportunities with Big Data 2011-1*. Cyber Center Technical Reports. http://docs.lib.purdue.edu/cctech/1.

Agrawal, D., S. Das, and A. El Abbadi. 2011. “Big Data and Cloud Computing: Current State and Future Opportunities.” In *Proceedings of the 14th International Conference on Extending Database Technology*, 530–533. ACM.

Aji, A., F. Wang, H. Vo, R. Lee, Q. Liu, X. Zhang, and J. Saltz. 2013. “Hadoop GIS: A High Performance Spatial Data Warehousing System Over Mapreduce.” *Proceedings of the VLDB Endowment* 6: 1009–1020.

Alam, S., F. D. Albareti, C. A. Prieto, F. Anders, S. F. Anderson, B. H. Andrews, E. Armengaud, et al. 2015. “The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III.” *The Astrophysical Journal Supplement Series* 219 (1): 1–27.

Alvaro, P., T. Condie, N. Conway, K. Elmeleegy, J. M. Hellerstein, and R. Sears. 2010. “Boom Analytics: Exploring Data-centric, Declarative Programming for the Cloud.” In *Proceedings of the 5th European Conference on Computer Systems*, 223–236. New York, NY: ACM.

Alyass, A., M. Turcotte, and D. Meyre. 2015. “From Big Data Analysis to Personalized Medicine for All: Challenges and Opportunities.” *BMC Medical Genomics* 8: 1–33.

Aminzadeh, N., Z. Sanaei, and S. H. Ab Hamid. 2015. “Mobile Storage Augmentation in Mobile Cloud Computing: Taxonomy, Approaches, and Open Issues.” *Simulation Modelling Practice and Theory* 50: 96–108.

Ammn, N., and M. Irfanuddin. 2013. “Big Data Challenges.” *International Journal of Advanced Trends in Computer Science and Engineering* 2 (1): 613–615.

Anderson, D. P., and G. Fedak. 2006. “The Computational and Storage Potential of Volunteer Computing.” In *Proceedings of the Sixth IEEE International Symposium on Cluster Computing and the Grid*, 73–80. Washington, DC: IEEE Computer Society.

Andreolini, M., M. Colajanni, M. Pietri, and S. Tosi. 2015. “Adaptive, Scalable and Reliable Monitoring of Big Data on Clouds.” *Journal of Parallel and Distributed Computing* 79–80 (C): 67–79.

Assunção, M. D., R. N. Calheiros, S. Bianchi, M. A. Netto, and R. Buyya. 2015. “Big Data Computing and Clouds: Trends and Future Directions.” *Journal of Parallel and Distributed Computing* 79–80: 3–15.

Aydin, G., I. R. Hallac, B. Karakus, G. Aydin, I. R. Hallac, and B. Karakus. 2015. “Architecture and Implementation of a Scalable Sensor Data Storage and Analysis System Using Cloud Computing and Big Data Technologies.” *Journal of Sensors, Journal of Sensors*. doi:10.1155/2015/834217.

Bae, B.-J., Y.-J. Kim, Y.-K. Kim, O.-K. Ha, and Y.-K. Jun. 2014. “An Intrusive Analyzer for Hadoop Systems Based on Wireless Sensor Networks.” *International Journal of Distributed Sensor Networks*. doi:10.1155/2014/196040.

Balakrishna, C. 2012. “Enabling Technologies for Smart City Services and Applications.” In *6th International Conference on Next Generation Mobile Applications, Services and Technologies (NGMAST2012)*, 223–227.

Banditwattanawong, T., M. Masdisornchote, and P. Uthayopas. 2014. “Economical and Efficient Big Data Sharing with I-Cloud.” In *International Conference on Big Data and Smart Computing (BIGCOMP), 2014*, 105–110.

Baughman, A. K., R. J. Bogdany, C. McAvoy, R. Locke, B. O’Connell, and C. Upton. 2015. “Predictive Cloud Computing with Big Data: Professional Golf and Tennis Forecasting [Application Notes].” *IEEE Computational Intelligence Magazine* 10 (3): 62–76.

Baumann, P., P. Mazzetti, J. Ungar, R. Barbera, D. Barboni, A. Beccati, L. Bigagli, et al. 2016. “Big Data Analytics for Earth Sciences: The EarthServer Approach.” *International Journal of Digital Earth* 9 (1): 3–29.

Belaud, J.-P., S. Negny, F. Dupros, D. Michéa, and B. Vautrin. 2014. “Collaborative Simulation and Scientific Big Data Analysis: Illustration for Sustainability in Natural Hazards Management and Chemical Process Engineering.” *Computers in Industry* 65 (3): 521–535.

Belissent, J. 2010. Getting Clever About Smart Cities: New Opportunities Require New Business Models. Accessed November 25, 2015. http://193.40.244.77/iot/wp-content/uploads/2014/02/getting_clever_about_smart_cities_new_opportunities.pdf.

Bellettini, C., M. Camilli, L. Capra, and M. Monga. 2013. “MaRDiGraS: Simplified Building of Reachability Graphs on Large Clusters.” *Reachability Problems* 8169: 83–95.

Benediktsson, J. A., J. Chanussot, and W. M. Moon. 2013. “Advances in Very-High-Resolution Remote Sensing.” *Proceedings of the IEEE* 101 (3): 566–569.

Benjamins, R., J. Contreras, O. Corcho, and A. Gomez-Perez. 2002. The Six Challenges of the Semantic Web. In *Proceedings of International Semantic Web Conference (ISWC2002)*, Sardinia, Italia, 2002. Accessed November 25, 2015. https://wikis.gsic.uva.es/juaase/images/b/bb/Benjaminsetal.pdf.

Berkovich, S., and D. Liao. 2012. On Clusterization of Big Data Streams. In *Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications*, 9. ACM. Accessed November 25, 2015. http://portalparts.acm.org/2350000/2345316/fm/frontmatter.pdf?ip = 100.36.182.180&CFID = 791206509&CFTOKEN = 86783680.

Bertino, E., P. Bernstein, D. Agrawal, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, et al. 2011. Challenges and Opportunities with Big Data. Accessed November 25, 2015. http://docs.lib.purdue.edu/ccpubs/445/.

Bertino, E., and M. Kantarcioglu. 2014. Big Data – Security with Privacy. Accessed November 25, 2015. https://www.cs.purdue.edu/homes/bertino/RFI-Response-NSF-BigData-SP-Oct16.pdf.

Bicer, T., D. Chiu, and G. Agrawal. 2012. Time and Cost Sensitive Data-Intensive Computing on Hybrid Clouds. In *Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing* (ccgrid 2012), 636–643. IEEE Computer Society.

Bird, I. 2011. “Computing for the Large Hadron Collider.” *Annual Review of Nuclear and Particle Science* 61: 99–118.

Bo, Y., and H. Wang. 2011. The Application of Cloud Computing and the Internet of Things in Agriculture and Forestry. In *International Joint Conference on Service Sciences (IJCSS), 2011*, 168–172.

Bonomi, F., R. Milito, J. Zhu, and S. Addepalli. 2012. “Fog Computing and its Role in the Internet of Things.” *Proceedings of the MCC Workshop on Mobile Cloud Computing*. doi:10.1145/2342509.2342513.

Boyd, D., and K. Crawford. 2012. “Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon.” *Information, Communication & Society* 15 (5): 662–679.

Brunswicker, S., E. Bertino, and S. Matei. 2015. “Big Data for Open Digital Innovation – A Research Roadmap.” *Big Data Research* 2 (2): 53–58.

Bryant, R., R. H. Katz, and E. D. Lazowska. 2008. Big-data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society. Accessed November 25, 2015. http://www.datascienceassn.org/sites/default/files/Big%20Data%20Computing%202008%20Paper.pdf.

Bughin, J., M. Chui, and J. Manyika. 2010. “Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch.” *McKinsey Quarterly* 56 (1): 75–86.

Burtica, R., E. M. Mocanu, M. I. Andreica, and N. Ţăpuş. 2012. Practical Application and Evaluation of No-SQL Databases in Cloud Computing. In *Proceedings of the 2012 IEEE International Systems Conference (SysCon)*, 1–6.

Camarinha-Matos, L. M., S. Tomic, and P. Graça, (Eds.). 2013. Technological Innovation for the Internet of Things: 4th IFIP WG 5.5/SOCOLNET. *Doctoral Conference on Computing, Electrical and Industrial Systems, DoCEIS 2013*, Costa de Caparica, Portugal, April 15–17, 2013, *Proceedings* (Vol. 394). Springer.

Campa, S., M. Danelutto, M. Goli, H. González-Vélez, A. M. Popescu, and M. Torquati. 2014. "Parallel Patterns for Heterogeneous CPU/GPU Architectures: Structured Parallelism from Cluster to Cloud." *Future Generation Computer Systems*, 37: 354–366.

Cao, G., S. Wang, M. Hwang, A. Padmanabhan, Z. Zhang, and K. Soltani. 2015. "A Scalable Framework for Spatiotemporal Analysis of Location-Based Social Media Data." *Computers, Environment and Urban Systems* 51: 70–82.

Cao, Y., C. Yang, and D. Wong. 2009. "An Interoperable Spatiotemporal Weather Radar Data Disseminating System." *International Journal of Remote Sensing* 30: 1313–1326.

Cary, A. 2011. Scaling Geospatial Searches in Large Spatial Databases. Accessed May 27, 2016. http://140.98.202.196/xpl/abstractReferences.jsp?reload = true&tp=&arnumber = 5576271&url = http%3A%2F%2F140.98.202.196%2Fxpls%2Ficp.jsp%3Farnumber%3D5576271.

Cary, A., Y. Yesha, M. Adjouadi, and N. Rishe. 2010. Leveraging Cloud Computing in Geodatabase Management. In *IEEE International Conference on Granular Computing (GrC)*, 73–78. IEEE.

Castiglione, A., M. Gribaudo, M. Iacono, and F. Palmieri. 2014. "Modeling Performances of Concurrent Big Data Applications." *Software: Practice and Experience* 45 (8): 1127–1144.

Cavoukian, A., and J. Jonas. 2012. Privacy by Design in the Age of Big Data. Information and Privacy Commissioner of Ontario, Canada. Accessed December 3 2015. https://privacybydesign.ca/content/uploads/2012/06/pbd-big_data.pdf.

Cheatham, M. 2015. Privacy in the Age of Big Data. In *The 2015 International Conference on Collaboration Technologies and Systems (CTS)*, 334–335.

Chen, J., Y. Chen, X. Du, C. Li, J. Lu, S. Zhao, and X. Zhou. 2013. "Big Data Challenge: A Data Management Perspective." *Frontiers of Computer Science* 7 (2): 157–164.

Chen, H., R. H. Chiang, and V. C. Storey. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact." *MIS Quarterly* 36 (4): 1165–1188.

Chen, M., S. Mao, Y. Zhang, and V. C. Leung. 2014a. *Chapter 1, Big Data: Related Technologies, Challenges and Future Prospects*. Heidelberg: Springer.

Chen, Q., L. Wang, and Z. Shang. 2008. MRGIS: A MapReduce-Enabled High Performance Workflow System for GIS, EScience, 2008. In *IEEE Fourth International Conference on eScience, 2008 (eScience'08)*, 646–651. IEEE.

Chen, Z. K., S. Q. Yang, S. Tan, H. Zhao, L. He, G. Zhang, and H. Y. Yang. 2014b. "The Data Allocation Strategy Based on Load in NoSQL Database." *Applied Mechanics and Materials* 513–517: 1464–1469.

Chen, B. Y., H. Yuan, Q. Li, S.-L. Shaw, W. H. Lam, and X. Chen. 2015. "Spatiotemporal Data Model for Network Time Geographic Analysis in the Era of Big Data." *International Journal of Geographical Information Science* 30 (6): 1041–1071.

Chen, C. P., and C.-Y. Zhang. 2014. "Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data." *Information Sciences* 275: 314–347.

Chen, D., and H. Zhao. 2012. Data Security and Privacy Protection Issues in Cloud Computing. In *2012 International Conference on Computer Science and Electronics Engineering (ICCSEE)*, Vol.1, 647–651. IEEE.

Cheng, H., C. Rong, K. Hwang, W. Wang, and Y. Li. 2015. "Secure Big Data Storage and Sharing Scheme for Cloud Tenants." *China Communications* 12 (6): 106–115.

Cheng, H., H. Yang, and C. Rong. 2012. Distributed Systems Combined with Advanced Network: Evolution, Applications and Challenges. In *2012 8th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM)*, 1–4. IEEE.

Choi, C., J. Choi, and P. Kim. 2014. "Ontology-Based Access Control Model for Security Policy Reasoning in Cloud Computing." *The Journal of Supercomputing* 67 (3): 711–722.

Christen, P. 2014. Privacy Aspects in Big Data Integration: Challenges and Opportunities. In *Proceedings of the First International Workshop on Privacy and Security of Big Data*, 1–1. ACM.

Chung, W.-C., C.-C. Chen, J.-M. Ho, C.-Y. Lin, W.-L. Hsu, Y.-C. Wang, D. T. Lee, F. Lai, C.-W. Huang and Y.-J. Chang. 2014. "CloudDOE: A User-Friendly Tool for Deploying Hadoop Clouds and Analyzing High-Throughput Sequencing Data with MapReduce." *PLoS ONE* 9 (6): e98146.

Church, P., A. Goscinski, and C. Lefèvre. 2015. "Exposing HPC and Sequential Applications as Services Through the Development and Deployment of a SaaS Cloud." *Future Generation Computer Systems* 43–44: 24–37.

Coppersmith, D. 1994. "The Data Encryption Standard (DES) and Its Strength Against Attacks." *IBM Journal of Research and Development* 38 (3): 243–250.

Coskun, V., B. Ozdenizci, and K. Ok. 2013. "A Survey on Near Field Communication (NFC) Technology." *Wireless Personal Communications* 71 (3): 2259–2294.

Cosulschi, M., A. Cuzzocrea, and R. De Virgilio. 2013. Implementing bfs-Based Traversals of rdf Graphs over Mapreduce Efficiently. In Cluster, Cloud and Grid Computing (CCGrid), 2013 13th, May. IEEE/ACM International Symposium on, 569–574). IEEE.

Cukier, K. 2010. Data, Data Everywhere, Economist. Accessed November 25, 2015. http://www.economist.com/node/15557443.

Cuzzocrea, A., G. Fortino, and O. Rana. 2013. Managing Data and Processes in Cloud-Enabled Large-Scale Sensor Networks: State-Of-The-Art and Future Research Directions. In *2013 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, 583–588. IEEE.

Dantan, J., Y. Pollet, and S. Taibi. 2013. The GOAL Approach-A Goal-Oriented Algebraic Language. In *Proceedings of the 8th International Conference on Evaluation of Novel Approaches to Software Engineering*, 173–180.

Das, M., and S. Parthasarathy. 2009. Anomaly detection and spatio-temporal analysis of global climate system. In *Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data*, 142–150. ACM, June.

Dasgupta, A. 2013. Big Data: The Future Is in Analytics, Geospatial World. Accessed May 27, 2016. http://www.geospatialworld.net/article/big-data-the-future-is-in-analytics.

Dean, J., and S. Ghemawat. 2008. "MapReduce: Simplified Data Processing on Large Clusters." *Communications of the ACM* 51 (1): 107–113.

Demirkan, H., and D. Delen. 2013. "Leveraging the Capabilities of Service-Oriented Decision Support Systems: Putting Analytics and Big Data in Cloud." *Decision Support Systems* 55 (1): 412–421.

Denning, D. E., and P. J. Denning. 1979. "Data Security." *ACM Computing Surveys (CSUR)* 11 (3): 227–249.

Dillon, M. 2015. Big Universe, Big Data, Astronomical Opportunity. Accessed November 25, 2015. http://www.theguardian.com/science/across-the-universe/2015/jun/25/big-universe-big-data-astronomical-opportunity.

Ding, J. M., Y. Jiang, Q. X. Wang, Y. L. Liu, and M. J. Li. 2013. "A Data Localization Algorithm for Distributing Column Storage System of Big Data." *Advanced Materials Research* 756–759: 3089–3093.

Dobre, C., and F. Xhafa. 2014. "Parallel Programming Paradigms and Frameworks in Big Data Era." *International Journal of Parallel Programming* 42 (5): 710–738.

Dong, X. H., and S. Divesh. 2015. "Big Data Integration." *Synthesis Lectures on Data Management* 7 (1): 1–198.

Doody, P., and A. Shields. 2012. Mining Network Relationships in the Internet of Things. In *Proceedings of the 2012 International Workshop on Self-aware Internet of Things*, 7–12. ACM.

Du, F., A. Zhu, and F. Qi. 2016. "Interactive Visual Cluster Detection in Large Geospatial Datasets Based on Dynamic Density Volume Visualization." *Geocarto International* 31, 597–611.

Duan, L., W. N. Street, and E. Xu. 2011. "Healthcare Information Systems: Data Mining Methods in the Creation of a Clinical Recommender System." *Enterprise Information Systems* 5 (2): 169–181.

Dubey, R., and A. Gunasekaran. 2015. "Education and Training for Successful Career in Big Data and Business Analytics." *Industrial and Commercial Training* 47 (4): 174–181.

EarthCube. 2014. EarthCube Enterprise Governance Draft Charter. Accessed November10 2015. http://workspace.earthcube.org/.

Edlich, S., S. Singh, and I. Pfennigstorf. 2013. Future Mobile Access for Open-Data Platforms and the BBC-DaaS System. In IS&T/SPIE Electronic Imaging, 866710–866710. International Society for Optics and Photonics.

Edwards, P. N. (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*, 518. Cambridge, MA: MIT Press.

Eisenstein, M. 2015. "Big Data: The Power of Petabytes." *Nature* 527 (7576): S2–S4.

Eldawy, A., and M. F. Mokbel. 2013. "A Demonstration of Spatialhadoop: An Efficient Mapreduce Framework for Spatial Data." *Proceedings of the VLDB Endowment* 6 (12): 1230–1233.

Evangelinos, C., and C. Hill. 2008. "Cloud Computing for Parallel Scientific HPC Applications: Feasibility of Running Coupled Atmosphere-Ocean Climate Models on Amazon's EC2." *Ratio* 2: 2–34.

Fan, J., and H. Liu. 2013. "Statistical Analysis of Big Data on Pharmacogenomics." *Advanced Drug Delivery Reviews.* 65 (7): 987–1000.

Fang, S., L. Da Xu, Y. Zhu, J. Ahati, H. Pei, J. Yan, and Z. Liu. 2014. "An Integrated System for Regional Environmental Monitoring and Management Based on Internet of Things." *IEEE Transactions on Industrial Informatics* 10 (2): 1596–1605.

Färber, F., S. K. Cha, J. Primsch, C. Bornhövd, S. Sigg, and W. Lehner. 2012. "SAP HANA Database: Data Management for Modern Business Applications." *ACM Sigmod Record* 40 (4): 45–51.

Feller, E., L. Ramakrishnan, and C. Morin. 2015. "Performance and Energy Efficiency of Big Data Applications in Cloud Environments: A Hadoop Case Study." *Journal of Parallel and Distributed Computing* 79–80: 80–89.

Feng, W. L., Y. C. Duan, M. X. Huang, L. F. Dong, X. Y. Zhou, and T. Hu. 2014. "A Research on Smart Tourism Service Mechanism Based on Context Awareness." *Applied Mechanics and Materials* 519–520: 752–758.

Feng, D. G., M. Zhang, Y. Zhang, and Z. Xu. 2011. "Study on Cloud Computing Security." *Journal of software* 22 (1): 71–83.

Fox, P., and J. Hendler. 2011. "Changing the Equation on Scientific Data Visualization." *Science, Washington* 331 (6018): 705–708.

Fu, Z., J. Shu, J. Wang, Y. Liu, and S. Lee. 2015. “Privacy-Preserving Smart Similarity Search Based on Simhash over Encrypted Data in Cloud Computing.” *Journal of Internet Technology* 16 (3): 453–460.

Gantz, J., and D. Reinsel. 2011. “Extracting Value from Chaos.” *IDC iView* 1142 (2011): 1–12.

Gantz, J., and D. Reinsel. 2012. “The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the far East.” *IDC iView: IDC Analyze the Future* 2007: 1–16.

Gao, S., L. Li, W. Li, K. Janowicz, and Y. Zhang. 2014. “ Constructing Gazetteers from Volunteered Big Geo-Data Based on Hadoop.” *Computers, Environment and Urban Systems*. doi:10.1016/j.compenvurbsys.2014.02.004.

García, S., J. Luengo, and F. Herrera. 2015. “ Data Preprocessing in Data Mining.” *Intelligent Systems Reference Library* 72. doi:10.1007/978-3-319-10247-4 (Chapter 6).

Geller, G. N., and W. Turner. 2007. The Model Web: A Concept for Ecological Forecasting. In *Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International*, 2469–2472. IEEE.

Goiri, Í, K. Le, T. D. Nguyen, J. Guitart, J. Torres, and R. Bianchini. 2012. GreenHadoop: Leveraging Green Energy in Data-Processing Frameworks. In *Proceedings of the 7th ACM European Conference on Computer Systems*, 57–70. ACM. http://www.cs.rutgers.edu/~ricardob/papers/eurosys12.pdf.

Gölzer, P., P. Cato, and M. Amberg. 2015. Data Processing Requirements of Industry 4.0-Use Cases for Big Data Applications. Data Processing. Accessed November 25, 2015. http://aisel.aisnet.org/cgi/viewcontent.cgi?article = 1060&context = ecis2015_rip.

Goodchild, M. F. 2007. “Citizens as Sensors: The World of Volunteered Geography.” *GeoJournal* 69 (4): 211–221.

Gopalkrishnan, V., D. Steier, H. Lewis, and J. Guszcza. 2012. Big Data, Big Business: Bridging the Gap. In *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, 7–11. ACM.

Griebel, L., H. U. Prokosch, F. Köpcke, D. Toddenroth, J. Christoph, I. Leb, I. Engel, and M. Sedlmayr. 2015. “A Scoping Review of Cloud Computing in Healthcare.” *BMC Medical Informatics and Decision Making* 15 (1): 1–16.

Grimmer, J. 2015. “We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together.” *PS: Political Science & Politics* 48 (1): 80–83.

Grolinger, K., M. Capretz, E. Mezghani, and E. Exposito. 2013. Knowledge as a Service Framework for Disaster Data Management. In *2013 IEEE 22nd International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, 6570634.

Gu, R., X. Yang, J. Yan, Y. Sun, B. Wang, C. Yuan, and Y. Huang. 2014. “SHadoop: Improving MapReduce Performance by Optimizing Job Execution Mechanism in Hadoop Clusters.” *Journal of Parallel and Distributed Computing* 74 (3): 2166–2179.

Gubbi, J., R. Buyya, S. Marusic, and M. Palaniswami. 2013. “Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions.” *Future Generation Computer Systems* 29 (7): 1645–1660.

Gui, Z., C. Yang, J. Xia, Q. Huang, K. Liu, Z. Li, M. Yu, M. Sun, N. Zhou, and B. Jin. 2014. “A Service Brokering and Recommendation Mechanism for Better Selecting Cloud Services.” *PLoS ONE* 9 (8): e105297.

Gui, Z., M. Yu, C. Yang, Y. Jiang, S. Chen, J. Xia, Q. Huang, et al. 2016. “ Developing Subdomain Allocation Algorithms Based on Spatial and Communicational Constraints to Accelerate Dust Storm Simulation.” *PloS One* 11 (4): e0152250.

Hammoud, M., and M. F. Sakr. 2011. Locality-Aware Reduce Task Scheduling for MapReduce. In *2011 IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom)*, 570–576. IEEE.

Han, J., E. Haihong, G. Le, and J. Du. 2011. Survey on NoSQL database. In *2011 6th International Conference on Pervasive Computing and Applications (ICPCA)*, 363–366. IEEE.

Han, Q., S. Liang, and H. Zhang. 2015. “Mobile Cloud Sensing, Big Data, and 5G Networks Make an Intelligent and Smart World.” *IEEE Network* 29 (2): 40–45.

Hartog, J., Z. Fadika, E. Dede, and M. Govindaraju. 2012. Configuring a MapReduce framework for dynamic and efficient energy adaptation. In *2012 IEEE 5th International Conference on Cloud Computing (CLOUD)*, 914–921. IEEE.

Hashem, I. A. T., I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan. 2015. “The Rise of “Big Data” on Cloud Computing: Review and Open Research Issues.” *Information Systems* 47: 98–115.

Heffner, J. 2014. “Predictive Policing.” *GEO: Geoconnexion Internal* 13 (7): 20–23.

Hellerstein, J. M. 2010. Datalog Redux: Experience and Conjecture. In Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 1–2. ACM.

Hong, Z., X. Tong, W. Cao, S. Jiang, P. Chen, and S. Liu. 2015. “Rapid Three-Dimensional Detection Approach for Building Damage due to Earthquakes by the use of Parallel Processing of Unmanned Aerial Vehicle Imagery.” *Journal of Applied Remote Sensing* 9: 097292–097292.

Hori, M., E. Kawashima, and T. Yamazaki. 2010. “Application of Cloud Computing to Agriculture and Prospects in Other Fields.” *Fujitsu Scientific & Technical Journal* 46 (4): 446–454.

Hsu, M. H. 2008. “A Personalized English Learning Recommender System for ESL Students.” *Expert Systems with Applications* 34 (1): 683–688.

Hsu, C.-H., K. D. Slagter, and Y.-C. Chung. 2015. “Locality and Loading Aware Virtual Machine Mapping Techniques for Optimizing Communications in MapReduce Applications.” *Future Generation Computer Systems* 53: 43–54.

Huang, Q., G. Cervone, D. Jing, and C. Chang. 2015. DisasterMapper: A CyberGIS Framework for Disaster Management Using Social Media Data. In *ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data* ACM, Seattle, WA, USA.

Huang, M. L., L. F. Lu, and X. Zhang. 2015. "Using Arced Axes in Parallel Coordinates Geometry for High Dimensional BigData Visual Analytics in Cloud Computing." *Computing* 97 (4): 425–437.

Huang, M., and R. Rust. 2013. "IT-Related Service: A Multidisciplinary Perspective." *Journal of Service Research* 16 (3): 251–258.

Huang, Q., and D. W. Wong. 2016. "Activity Patterns, Socioeconomic Status and Urban Spatial Structure: What Can Social Media Data Tell Us?" *International Journal of Geographical Information Science* 30 (9): 1873–1898.

Huang, Q., and C. Xu. 2014. "A Data-Driven Framework for Archiving and Exploring Social Media Data." *Annals of GIS* 20: 265–277.

Huang, Q., and C. Yang. 2011. "Optimizing Grid Computing Configuration and Scheduling for Geospatial Analysis: An Example with Interpolating DEM." *Computers & Geosciences* 37 (2): 165–176.

Huang, Q., C. Yang, K. Benedict, S. Chen, A. Rezgui, and J. Xie. 2013a. "Utilize Cloud Computing to Support Dust Storm Forecasting." *International Journal of Digital Earth* 6: 338–355.

Huang, Q., C. Yang, K. Benedict, J. Xie, A. Rezgui, J. Xia, and S. Chen. 2013b. "Using Spatiotemporal Patterns and High end Computing to Enable Dust Storm Forecasting." *International Journal of Geographical Information Science* 27 (4): 765–784.

Hung, P. P., M. Aazam, and E.-N. Huh. 2015. "CTaG: An Innovative Approach for Optimizing Recovery Time in Cloud Environment." *Transactions on Internet and Information Systems* 9 (4): 1282–1301.

Hung, P. P., B. Tuan-Anh, and E.-N. Huh. 2013. A Solution of Thin-thick Client Collaboration for data Distribution and Resource Allocation in Cloud Computing. In *2013 International Conference on Information Networking (ICOIN)*, 238–243. IEEE.

Indeck, R. S., and D. M. Indeck. 2012. *U.S. Patent No. 8,156,101*. Washington, DC: U.S. Patent and Trademark Office.

Internet Live Stats. 2016. Accessed 27 September 2016. http://www.internetlivestats.com/internet-users/

Ismail, L., and L. Khan. 2015. "Implementation and Performance Evaluation of a Scheduling Algorithm for Divisible Load Parallel Applications in a Cloud Computing Environment." *Software: Practice and Experience* 45 (6): 765–781.

Jagadish, H. V., J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi. 2014. "Big Data and its Technical Challenges." *Communications of the ACM* 57 (7): 86–94.

Jalali, A., O. A. Olabode, and C. M. Bell. 2012. "Leveraging Cloud Computing to Address Public Health Disparities: An Analysis of the SPHPS." *Online Journal of Public Health Informatics* 4 (3: 1–7).

Jam, M. R., L. M. Khanli, M. K. Akbari, E. Hormozi, and M. S. Javan. 2013. Survey on Improved Autoscaling in Hadoop into Cloud Environments. In *2013 5th Conference on Information and Knowledge Technology (IKT)*, 19–23. IEEE.

Jayalath, C., J. Stephen, and P. Eugster. 2014. "From the Cloud to the Atmosphere: Running MapReduce Across Data Centers." *IEEE Transactions on Computers* 63 (1): 74–87.

Ji, C., Y. Li, W. Qiu, U. Awada, and K. Li. 2012. Big Data Processing in Cloud Computing Environments. In *2012 12th International Symposium on Pervasive Systems, Algorithms and Networks (ISPAN)*, 17–23. IEEE.

Jiang, H., Y. Chen, Z. Qiao, T. H. Weng, and K. C. Li. 2015. "Scaling up Mapreduce-Based Big Data Processing on Multi-GPU Systems." *Cluster Computing* 18 (1): 369–383.

Jiang, S., L. Fang, and X. Huang. 2009. An Idea of Special Cloud Computing in Forest Pests' Control. In *Cloud Computing*, 615–620. Springer Berlin Heidelberg.

Jiang, Y., Y. Li, C. Yang, E. M. Armstrong, T. Huang, and D. Moroni. 2016. "Reconstructing Sessions From Data Discovery and Access Logs to Build a Semantic Knowledge Base for Improving Data Discovery." *ISPRS International Journal of Geo-Information* 5 (4): 54–68.

Jin, J., J. Gubbi, S. Marusic, and M. Palaniswami. 2014. "An Information Framework for Creating a Smart City Through Internet of Things." *IEEE Internet of Things Journal* 1 (2): 112–121.

Joo, H., B. Hong, and S. Kim. 2012. Smart-Contents Visualization of Publishing Big Data Using NFC Technology. In *Computer Applications for Graphics, Grid Computing, and Industrial Environment*, 118–123. Springer.

Kagermann, H., J. Helbig, A. Hellinger, and W. Wahlster. 2013. *Recommendations for Implementing the Strategic Initiative INDUSTRIE 4.0: Securing the Future of German Manufacturing Industry.* Final Report of the Industrie 4.0 Working Group. Forschungsunion.

Kaisler, S., F. Armour, J. A. Espinosa, and W. Money. 2013. Big Data: Issues and Challenges Moving Forward. In *2013 46th Hawaii International Conference on System Sciences (HICSS)*, 995–1004.

Karimi, H. A. ed. 2014. *Big Data: Techniques and Technologies in Geoinformatics*. Boca Raton, FL: CRC Press.

Kaufman, L. M. 2009. "Data Security in the World of Cloud Computing." *IEEE Security & Privacy Magazine* 7 (4): 61–64.

Kaushik, R. T., and M. Bhandarkar. 2010. Greenhdfs: Towards an Energy-Conserving, Storage-Efficient, Hybrid Hadoop Compute Cluster. In *Proceedings of the USENIX Annual Technical Conference*, 109.

Kaushik, R. T., M. Bhandarkar, and K. Nahrstedt. 2010. Evaluation and Analysis of Greenhdfs: A self-adaptive, Energy-Conserving Variant of the Hadoop Distributed File System. In *2010 IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom)*, 274–287.

Kehoe, B., S. Patil, P. Abbeel, and K. Goldberg. 2015. "A Survey of Research on Cloud Robotics and Automation." *IEEE Transactions on Automation Science and Engineering* 12 (2): 398–409.

Khan, N., I. Yaqoob, I. A. T. Hashem, Z. Inayat, W. K. Mahmoud Ali, M. Alam, M. Shiraz, and A. Gani. 2014. "Big Data: Survey, Technologies, Opportunities, and Challenges." *The Scientific World Journal* 2014: 1–18.

Kim, C. 2014. "Theoretical Analysis of Constructing Wavelet Synopsis on Partitioned Data Sets." *Multimedia Tools and Applications* 74 (7): 2417–2432.

Kim, S., D. Kang, J. Choi, and J. Kim. 2014. Burstiness-aware I/O scheduler for MapReduce framework on virtualized environments. In *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*, 305–308.

Kim, W., H. Kim, and Y. Kim. 2013. "DataConnector: A Data Processing Framework Integrating Hadoop and aGrid Middleware OGSA-DAI for Cloud Environment." *Information – An International Interdisciplinary Journal.* 16 (1B): 801–806.

Kim, G. H., S. Trimi, and J. H. Chung. 2014. "Big-data Applications in the Government Sector." *Communications of the ACM* 57 (3): 78–85.

Kim, I.-H., M.-H. Tsou, and C.-C. Feng. 2015. "Design and Implementation Strategy of a Parallel Agent-Based Schelling Model." *Computers, Environment and Urban Systems* 49, 30–41.

Knoppers, B. M., and R. Chadwick. 2014. "Human Genetic Research: Emerging Trends in Ethics." *Focus* 4 (3): 416–422.

Kos, A., S. Tomažič, J. Salom, N. Trifunovic, M. Valero, and V. Milutinovic. 2015. "New Benchmarking Methodology and Programming Model for Big Data Processing." *International Journal of Distributed Sensor Networks* 501: 1–7.

Kourtesis, D., J. M. Alvarez-Rodríguez, and I. Paraskakis. 2014. "Semantic-based QoS Management in Cloud Systems: Current Status and Future Challenges." *Future Generation Computer Systems* 32: 307–323.

Kozuch, M. A., M. P. Ryan, R. Gass, S. W. Schlosser, D. O'Hallaron, J. Cipar, E. Krevat, J. López, M. Stroucken and G. R. Ganger. 2009. Tashi: Location-aware Cluster Management. In *Proceedings of the 1st Workshop on Automated Control for Datacenters and Clouds*, 43–48. ACM.

Krämer, M., and I. Senner. 2015. "A Modular Software Architecture for Processing of big Geospatial Data in the Cloud." *Computers & Graphics* 49: 69–81.

Krogh, B. H. 2008. *Cyber Physical Systems: The Need for New Models and Design Paradigms*. Presentation Report. Accessed November 30, 2015. http://slideplayer.com/slide/4807731/.

Ku, M., E. Choi, and D. Min. 2014. "An Analysis of Performance Factors on Esper-Based Stream big Data Processing in a Virtualized Environment." *International Journal of Communication Systems* 27 (6): 898–917.

Lampos, V., and N. Cristianini. 2010. Tracking the Flu Pandemic by Monitoring the Social Web. In *2010 2nd International Workshop on Cognitive Information Processing (CIP)*, 411–416. IEEE.

Lary, D. J., S. Woolf, F. Faruque, and J. P. LePage. 2014. "Holistics 3.0 for Health." *ISPRS International Journal of Geo-Information* 3: 1023–1038.

LaValle, S., E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz. 2013. Big data, Analytics and the Path from Insights to Value. *MIT Sloan Management Review*, 21. Accessed December 2, 2015. http://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/.

Lee, S., H. Park, and Y. Shin. 2012. "Cloud Computing Availability: Multi-Clouds for big Data Service." *Convergence and Hybrid Information Technology* 310: 799–806.

Li, Z. 2015. Optimizing Geospatial Cyberinfrastructure to Improve the Computing Capability for Climate Studies. Accessed November 25, 2015. http://digilib.gmu.edu/jspui/bitstream/handle/1920/9630/Li_gmu_0883E_10873.pdf?sequence = 1&isAllowed = y.

Li, Z., F. Hu, J. Schnase, D. Duffy, T. Lee, C. Yang, and M. Bowen. 2016a. "A Spatiotemporal Indexing Approach for Efficient Process of Big Array-Based Climate Data with MapReduce." *International Journal of Geographic Information Science*. doi:10.1080/13658816.2015.1131830.

Li, W., M. Song, B. Zhou, K. Cao, and S. Gao. 2015a. "Performance Improvement Techniques for Geospatial web Services in a Cyberinfrastructure Environment–A Case Study with a Disaster Management Portal." *Computers, Environment and Urban Systems* 54: 314–325.

Li, Z., C. Yang, Q. Huang, K. Liu, M. Sun, and J. Xia. 2014. "Building Model as a Service for Supporting Geosciences." *Computers, Environment and Urban Systems*. doi:10.1016/j.compenvurbsys.2014.06.004.

Li, Z., C. Yang, K. Liu, H. Fei, and B. Jin. 2016b. "Automatic Scaling Hadoop in the Cloud for Efficient Process of Big Geospatial Data." *ISPRS International Journal of Geo-Information* 5 (10): 173. doi:10.3390/ijgi5100173.

Li, Z., C. Yang, H. Wu, W. Li, and L. Miao. 2011. "An Optimized Framework for Seamlessly Integrating OGC Web Services to Support Geospatial Sciences." *International Journal of Geographical Information Science* 25 (4): 595–613.

Li, Z., C. Yang, M. Yu, K. Liu, and M. Sun. 2015b. "Enabling Big Geoscience Data Analytics with a Cloud-Based, MapReduce-Enabled and Service-Oriented Workflow Framework." *PloS one* 10 (3): e0116781.

Lim, N., Å Grönlund, and A. Andersson. 2015. "Cloud Computing: The Beliefs and Perceptions of Swedish School Principals." *Computers & Education* 84: 90–100.

Lin, F. C., L. K. Chung, W. Y. Ku, L. R. Chu, and T. Y. Chou. 2013. The Framework of Cloud Computing Platform for Massive Remote Sensing Images. In *2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA)*, 621–628. IEEE.

Liu, L. 2013. "Computing Infrastructure for big Data Processing." *Frontiers of Computer Science* 7 (2): 165–170.

Liu, X. J. 2014. "Research of Big Data Processing Platform." *In Applied Mechanics and Materials* 484: 922–926.

Liu, Z. Z., Z. P. Jia, X. Xue, and J. Y. An. 2015. "Reliable Web Service Composition Based on QoS Dynamic Prediction." *Soft Computing* 19 (5): 1409–1425.

Liu, X., C. Kang, L. Gong, and Y. Liu. 2016. "Incorporating Spatial Interaction Patterns in Classifying and Understanding Urban Land use." *International Journal of Geographical Information Science* 30, 334–350.

Liu, K., C. Yang, W. Li, Z. Gui, C. Xu, and J. Xia. 2014. "Using Semantic Search and Knowledge Reasoning to Improve the Discovery of Earth Science Records: An Example with the ESIP Semantic Testbed." *International Journal of Applied Geospatial Research (IJAGR)* 5 (2): 44–58.

Liu, K., C. Yang, W. Li, Z. Li, H. Wu, A. Rezgui, and J. Xia. 2011. The GEOSS Clearinghouse High Performance Search Engine. In *2011 19th International Conference on Geoinformatics*, 1–4.

Lohr, S. 2012. The age of big data. *New York Times*, 11.

Lorido-Botrán, T., J. Miguel-Alonso, and J. A. Lozano. 2012. *Auto-scaling Techniques for Elastic Applications in Cloud Environments*. Department of Computer Architecture and Technology, University of Basque Country, Tech. Rep. EHU-KAT-IK-09, 12, 2012.

Luo, C., J. Zhan, Z. Jia, L. Wang, G. Lu, L. Zhang, C.-Z. Xu, and N. Sun. 2012. "Cloudrank-d: Benchmarking and Ranking Cloud Computing Systems for Data Processing Applications." *Frontiers of Computer Science* 6 (4): 347–362.

Lushbough, C. M., E. Z. Gnimpieba, and R. Dooley. 2015. "Life Science Data Analysis Workflow Development Using the Bioextract Server Leveraging the IPlant Collaborative Cyberinfrastructure." *Concurrency and Computation: Practice and Experience* 27 (2): 408–419.

Ma, X., Y. Wang, X. Pei, and F. Xu. 2015. "Scalable and Elastic Total Order in Content-Based Publish/Subscribe Systems." *Computer Networks* 83: 297–314. doi:10.1016/j.comnet.2015.04.001.

Manuel, P. 2013. "A Trust Model of Cloud Computing Based on Quality of Service." *Annals of Operations Research* 233: 1–12.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. 2011. Big data: The Next Frontier for Innovation, Competition, and Productivity. Accessed November 25, 2015. http://www.citeulike.org/group/18242/article/9341321.

Mao, M., and M. Humphrey. 2011. Auto-scaling to Minimize Cost and Meet Application Deadlines in Cloud Workflows. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, 49. ACM.

Marr, B. 2015. *Big Data: Using SMART Big Data. Analytics and Metrics To Make Better Decisions and Improve Performance.* Atrium: Wiley.

Massey, N., R. Jones, F. E. L. Otto, T. Aina, S. Wilson, J. M. Murphy, D. Hassell, Y. H. Yamazaki, and M. R. Allen. 2014. "weather@home-development and Validation of a Very Large Ensemble Modelling System for Probabilistic Event Attribution." *Quarterly Journal of the Royal Meteorological Society* 141 (690): 1528–1545.

Mayer-Schönberger, V., and K. Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt. (Chapter 1).

McAfee, A., E. Brynjolfsson, T. H. Davenport, D. J. Patil, and D. Barton. 2012. "Big Data. The Management Revolution." *Harvard Business Review* 90 (10): 61–67.

Megler, V. M., and D. Maier. 2012. When Big Data Leads to Lost Data. In *Proceedings of the 5th Ph. D. Workshop on Information and Knowledge*, 1–8. ACM.

Mell, P. M., and T. Grance. 2011. *The NIST Definition of Cloud Computing*. Special Publication 800-145, National Institute of Standards and Technology, Gaithersburg, MD. doi:10.6028/nist.sp.800-145.

Michael, K., and K. Miller. 2013. "Big Data: New Opportunities and New Challenges [Guest Editors" Introduction]." *Computer* 46 (6): 22–24.

Mitton, N., S. Papavassiliou, A. Puliafito, and K. S. Trivedi. 2012. "Combining Cloud and Sensors in a Smart City Environment." *EURASIP journal on Wireless Communications and Networking* 2012 (1): 247–10.

Miyano, T., and M. Uehara. 2012. Proposal for Cloud Search Engine as a Service. In *2012 15th International Conference on Network-Based Information Systems (NBiS)*, 627–632. IEEE.

Moniruzzaman, A. B. M., and S. A. Hossain. 2013. "Nosql Database: New era of Databases for big Data Analytics-Classification, Characteristics and Comparison." *International Journal of Database Theory and Application.* 6 (4): 1–14.

Moore, R., T. A. Prince, and M. Ellisman. 1998. "Data-intensive Computing and Digital Libraries." *Communications of the ACM* 41 (11): 56–62.

Nadeem, A., and M. Y. Javed. 2005. A Performance Comparison of Data Encryption Algorithms. In *ICICT 2005. First International Conference on Information and Communication Technologies, 2005*, 84–89.

Najjar, M. S., and W. J. Kettinger. 2013. “Data Monetization: Lessons from a Retailer’s Journey.” *MIS Quarterly Executive* 12 (4): 1–13.

Nambiar, R., R. Chitor, and A. Joshi. 2014. *Data Management – A Look Back and a Look Ahead, Specifying Big Data Benchmarks*. Toronto: Springer.

Nasser, T., and R. S. Tariq. 2015. “Big Data Challenges.” *Journal of Computer Engineering & Information Technology* 4 (3): 1–10.

Nativi, S., P. Mazzetti, M. Santoro, F. Papeschi, M. Craglia, and O. Ochiai. 2015. “Big Data Challenges in Building the Global Earth Observation System of Systems.” *Environmental Modelling & Software* 68, 1–26.

Nazir, A., Y. M. Yassin, C. P. Kit, and E. K. Karuppiah. 2012. Evaluation of Virtual Machine Scalability on Distributed Multi/many-core Processors for Big Data Analytics. In *2012 IEEE Conference on Open Systems (ICOS)*, 1–6. IEEE.

Nguyen, P., and M. Halem. 2011. A Mapreduce Workflow System for Architecting Scientific Data Intensive Applications. In *Proceedings of the 2nd International Workshop on Software Engineering for Cloud Computing*, 57–63. ACM.

Nita, M. C., C. Chilipirea, C. Dobre, and F. Pop. 2013. A SLA-based Method for Big-data Transfers with Multi-criteria Optimization Constraints for IaaS. In *2013 11th Roedunet International Conference (RoEduNet)*, 1–6.

O’Donovan, P., K. Leahy, K. Bruton, and D. T. O’Sullivan. 2015. “Big Data in Manufacturing: A systematic Mapping Study.” *Journal of Big Data* 2 (1): 1–22.

O’Driscoll, A., V. Belogrudov, J. Carroll, K. Kropp, P. Walsh, P. Ghazal, and R. D. Sleator. 2015. “HBLAST: Parallelised Sequence Similarity–A Hadoop MapReducable Basic Local Alignment Search Tool.” *Journal of Biomedical Informatics* 54: 58–64.

Oguntimilehin, A., and E. O. Ademola. 2014. “A Review of Big Data Management, Benefits and Challenges.” *A Review of Big Data Management, Benefits and Challenges* 5 (6): 433–438.

Padgavankar, M. H., and S. R. Gupta. 2014. “Big Data Storage and Challenges.” *(IJCSIT) International Journal of Computer Science and Information Technologies* 5 (2): 2218–2223.

Padhy, R. P., M. R. Patra, and S. C. Satapathy. 2011. “RDBMS to NoSQL: Reviewing Some Next-Generation non-Relational Databases.” *International Journal of Advanced Engineering Science and Technologies* 11 (1): 15–30.

Padmanabhan, A., S. Wang, G. Cao, M. Hwang, Y. Zhao, Z. Zhang, and Y. Gao. 2013. FluMapper: An interactive CyberGIS Environment for Massive Location-Based Social Media Data Analysis. In *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*, 33. ACM.

Pandey, S., and S. Nepal. 2013. “Cloud Computing and Scientific Applications – Big Data, Scalable Analytics, and Beyond.” *Future Generation Computer Systems* 29: 1774–1776.

Paquette, J., and T. Tokuyasu. 2011. Hypergraph Visualization and Enrichment Statistics: How the EGAN Paradigm Facilitates Organic Discovery from Big Data. In *IS&T/SPIE Electronic Imaging*, 78650E–78650E. International Society for Optics and Photonics.

Perera, C., R. Ranjan, L. Wang, S. U. Khan, and A. Y. Zomaya. 2015. “Big Data Privacy in the Internet of Things Era.” *IT Professional* 17 (3): 32–39.

Pijanowski, B. C., A. Tayyebi, J. Doucette, B. K. Pekin, D. Braun, and J. Plourde. 2014. “A big Data Urban Growth Simulation at a National Scale: Configuring the GIS and Neural Network Based Land Transformation Model to run in a High Performance Computing (HPC) Environment.” *Environmental Modelling & Software* 51, 250–268.

Pokorny, J. 2013. “NoSQL Databases: A Step to Database Scalability in web Environment.” *International Journal of Web Information Systems* 9 (1): 69–82.

Pop, F., C. Dobre, V. Cristea, N. Bessis, F. Xhafa, and L. Barolli. 2015. “Deadline Scheduling for Aperiodic Tasks in Inter-Cloud Environments: A new Approach to Resource Management.” *The Journal of Supercomputing* 71 (5): 1754–1765.

Pumma, S., T. Achalakul, and X. Li. 2012. Automatic VM Allocation for Scientific Application. In *2012 IEEE 18th International Conference on Parallel and Distributed Systems (ICPADS)*, 828–833. IEEE.

Radke, A. M., and M. M. Tseng. 2015. “Design Considerations for Building Distributed Supply Chain Management Systems Based on Cloud Computing.” *Journal of Manufacturing Science and Engineering* 137 (4): 1–7.

Ramapriyan, H. K. 2015. The Role and Evolution of NASA’s Earth Science Data Systems. Accessed November 25, 2015. http://ntrs.nasa.gov/search.jsp?R = 20150018076.

Rasmussen, A., M. Conley, G. Porter, R. Kapoor, and A. Vahdat. 2012. Themis: An I/O-efficient MapReduce. In *Proceedings of the Third ACM Symposium on Cloud Computing*, 13. ACM.

Reda, K., A. Febretti, A. Knoll, J. Aurisano, J. Leigh, A. Johnson, M. E. Papka, and M. Hereld. 2013. “Visualizing Large, Heterogeneous Data in Hybrid-Reality Environments.” *IEEE Computer Graphics and Applications* 33(4): 38–48.

Redlich, R. M., and M. A. Nemzow. 2006. *U.S. Patent No. 7,103,915*. Washington, DC: U.S. Patent and Trademark Office.

Robinson, S. 2012. The Storage and Transfer Challenges of Big Data. Accessed November 25, 2015. http://sloanreview.mit.edu/article/the-storage-and-transfer-challenges-of-big-data/.

Röme, T. 2010. “Autoscaling Hadoop Clusters.” MSc thesis, University of Tartu. Accessed November 25, 2015. http://lepo.it.da.ut.ee/~srirama/publications/theses/AutoscaleHadoop_Toomas.pdf.

Sagl, G., M. Loidl, and E. Beinat. 2012. “A Visual Analytics Approach for Extracting Spatio-Temporal Urban Mobility Information from Mobile Network Traffic.” *ISPRS International Journal of Geo-Information* 1: 256–271.

Sainio, J., J. Westerholm, and J. Oksanen. 2015. “Generating Heat Maps of Popular Routes Online from Massive Mobile Sports Tracking Application Data in Milliseconds While Respecting Privacy.” *ISPRS International Journal of Geo-Information* 4: 1813–1826.

Sandhu, R., and S. K. Sood. 2015a. “A Commercial, Benefit Driven and Secure Framework for Elearning in Cloud Computing.” *Computer Applications in Engineering Education* 23 (4): 499–513.

Sandhu, R., and S. K. Sood. 2015b. “Scheduling of Big Data Applications on Distributed Cloud Based on QoS Parameters.” *Cluster Computing* 18 (2): 817–828.

SAS. 2012. *Data Visualization: Making Big Data Approachable and Valuable*. White Paper. A Survey on Information Visualization: Recent Advances and Challenges. Accessed November 25, 2015. https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/sas-data-visualization-marketpulse-106176.pdf.

Schaffers, H., N. Komninos, M. Pallot, B. Trousse, M. Nilsson, and A. Oliveira. 2011. “Smart Cities and the Future Internet: Towards Cooperation Frameworks for Open Innovation.” *Future Internet Assembly* 6656: 431–446.

Schlieski, T., and B. D. Johnson. 2012. “Entertainment in the age of big Data.” *Proceedings of the IEEE* 100: 1404–1408.

Schnase, J. L., D. Q. Duffy, G. S. Tamkin, D. Nadeau, J. H. Thompson, et al. 2014. “MERRA Analytic Services: Meeting the big Data Challenges of Climate Science Through Cloud-Enabled Climate Analytics-as-A-Service.” *Computers, Environment and Urban Systems*. doi:10.1016/j.compenvurbsys.2013.12.003.

Sequeira, H., P. Carreira, T. Goldschmidt, and P. Vorst. 2014. Energy Cloud: Real-time Cloud-Native Energy Management System to Monitor and Analyze Energy Consumption in Multiple Industrial Sites. In *Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, 529–534. IEEE Computer Society.

Sfrent, A., and F. Pop. 2015. “Asymptotic Scheduling for Many Task Computing in Big Data Platforms.” *Information Sciences* 319: 71–91.

Shekhar, S., V. Gunturi, M. R. Evans, and K. Yang. 2012. Spatial Big-data Challenges Intersecting Mobility and Cloud Computing. In *Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, 1–6. ACM.

Shelton, T., A. Poorthuis, and M. Zook. 2015. “Social Media and the City: Rethinking Urban Socio-Spatial Inequality Using User-Generated Geographic Information.” *Landscape and Urban Planning* 142: 198–211.

Shen, Z., S. Subbiah, X. Gu, and J. Wilkes. 2011. Cloudscale: Elastic Resource Scaling for Multi-tenant Cloud Systems. In *Proceedings of the 2nd ACM Symposium on Cloud Computing*, 5. ACM.

Shook, E., M. E. Hodgson, S. Wang, B. Behzad, K. Soltani, A. Hiscox, and J. Ajayakumar. 2016. “Parallel Cartographic Modeling: A Methodology for Parallelizing Spatial Data Processing.” *International Journal of Geographical Information Science* 30: 2355–2376.

Shvachko, K., H. Kuang, S. Radia, and R. Chansler. 2010. The Hadoop Distributed File System. In *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, 1–10. IEEE.

Singh, G., S. Bharathi, A. Chervenak, E. Deelman, C. Kesselman, M. Manohar, S. Patil, and L. Pearlman. 2003. A Metadata Catalog Service for Data Intensive Applications. In *Supercomputing, 2003 ACM/IEEE Conference*, 33–33. IEEE.

Slagter, K., C. H. Hsu, and Y. C. Chung. 2015. “An Adaptive and Memory Efficient Sampling Mechanism for Partitioning in MapReduce.” *International Journal of Parallel Programming* 43 (3): 489–507.

Slagter, K., C.-H. Hsu, Y.-C. Chung, and D. Zhang. 2013. “An Improved Partitioning Mechanism for Optimizing Massive Data Analysis Using MapReduce.” *The Journal of Supercomputing* 66 (1): 539–555.

Smid, M. E., and D. K. Branstad. 1988. “Data Encryption Standard: Past and Future.” *Proceedings of the IEEE* 76 (5): 550–559.

Somasundaram, T. S., K. Govindarajan, V. Venkateswaran, R. Radhika, and V. Venkatesh. 2012. CDM Server: A Data Management Framework for Data Intensive Application in Internal Private Cloud Infrastructure. In *2012 Seventh International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)*, 211–217. IEEE.

Soyata, T., R. Muraleedharan, J. Langdon, C. Funai, S. Ames, M. Kwon, and W. Heinzelman. 2012. COMBAT: Mobile-Cloud-based cOmpute/coMmunications Infrastructure for BATtlefield applications. In *SPIE Defense, Security, and Sensing*, 84030K–84030K. International Society for Optics and Photonics.

Sultan, N. 2010. “Cloud Computing for Education: A new Dawn?” *International Journal of Information Management* 30 (2): 109–116.

Sun, D. W., G. R. Chang, D. Chen, and X. W. Wang. 2014. “Profiling, Quantifying, Modeling and Evaluating Green Service Level Objectives in Cloud Computing Environments.” *Chinese Journal Of Computers* 36 (7): 1509–1525.

Sun, M., J. Li, C. Yang, G. A. Schmidt, M. Bambacus, R. Cahalan, Q. Huang, C. Xu, E. U. Noble, and Z. Li. 2012. “A Web-Based Geovisual Analytical System for Climate Studies.” *Future Internet* 4: 1069–1085.

Sun, D., G. Zhang, S. Yang, W. Zheng, S. U. Khan, and K. Li. 2015. “Re-Stream: Real-Time and Energy-Efficient Resource Scheduling in big Data Stream Computing Environments.” *Information Sciences* 319: 92–112. doi:10.1016/j.ins.2015.03.027.

Tablan, V., I. Roberts, H. Cunningham, and K. Bontcheva. 2013. “GATECloud.net: A Platform for Large-Scale, Open-Source Text Processing on the Cloud.” *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 371 (1983): 1–13.

Tamura, Y., and S. Yamada. 2015. “Reliability Analysis Based on a Jump Diffusion Model with Two Wiener Processes for Cloud Computing with Big Data.” *Entropy* 17 (7): 4533–4546.

Tang, Z., L. Jiang, J. Zhou, K. Li, and K. Li. 2015. “A Self-Adaptive Scheduling Algorithm for Reduce Start Time.” *Future Generation Computer Systems* 43–44: 51–60.

Tene, O. 2011. “Privacy: The new Generations.” *International Data Privacy Law* 1 (1): 15–27.

Terry, N. 2012. Protecting Patient Privacy in the Age of Big Data. Accessed November 25, 2015. http://papers.ssrn.com/sol3/papers.cfm?abstract_id = 2153269.

Terry, D. B., V. Prabhakaran, R. Kotla, M. Balakrishnan, M. K. Aguilera, and H. Abu-Libdeh. 2013. Consistency-Based Service Level Agreements for Cloud Storage. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, 309–324. ACM.

Theodoridis, E., G. Mylonas, and I. Chatzigiannakis. 2013. Developing an IoT Smart City framework. In *Information, Intelligence, Systems and Applications (IISA) 2013*. doi:10.1109/iisa.2013.6623710.

The Whitehouse. 2014. Harnessing Observations and Data about Our Earth. Accessed November 25, 2015. https://www.whitehouse.gov/blog/2014/07/18/harnessing-observations-and-data-about-our-earth.

Tien, J. M. 2013. “Big Data: Unleashing Information.” *Journal of Systems Science and Systems Engineering* 22 (2): 127–151.

Toole, J. L., S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González. 2015. “The Path Most Traveled: Travel Demand Estimation Using big Data Resources.” *Transportation Research Part C: Emerging Technologies* 58: 162–177.

Triguero, I., D. Peralta, J. Bacardit, S. García, and F. Herrera. 2015. “MRPR: A MapReduce Solution for Prototype Reduction in big Data Classification.” *Neurocomputing* 150: 331–345.

Van den Dam, R. 2013. Internet of Things: The Foundational Infrastructure for a Smarter Planet. In *Internet of Things, Smart Spaces, and Next Generation Networking*, 1–12. Springer Berlin Heidelberg. (Chapter 1).

Van Der Aalst, W. M., A. H. Ter Hofstede, B. Kiepuszewski, and A. P. Barros. 2003. “Workflow Patterns.” *Distributed and Parallel Databases* 14 (1): 5–51.

Van Der Aalst, W., and K. M. Van Hee. 2004. *Workflow Management: Models, Methods, and Systems*. Cambridge, MA: MIT press.

Vasile, M. A., F. Pop, R. I. Tutueanu, V. Cristea, and J. Kołodziej. 2014. “Resource-aware Hybrid Scheduling Algorithm in Heterogeneous Distributed Computing.” *Future Generation Computer Systems* 51: 61–71. doi:10.1016/j.future.2014.11.019.

Vatsavai, R. R., A. Ganguly, V. Chandola, A. Stefanidis, S. Klasky, and S. Shekhar. 2012. Spatiotemporal Data Mining in the Era of Big Spatial Data: Algorithms and Applications. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, 1–10. ACM.

Villars, R. L., C. W. Olofson, and M. Eastwood. 2011. *Big Data: What It Is and Why You Should Care*. White Paper, IDC.

Wang, M., S. B. Handurukande, and M. Nassar. 2012. RPig: A Scalable Framework for Machine Learning and Advanced Statistical Functionalities. In *IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom)*, December 2012, 293–300. IEEE.

Wang, Y., Z. Liu, H. Liao, and C. Li. 2015. “Improving the Performance of GIS Polygon Overlay Computation with MapReduce for Spatial big Data Processing.” *Cluster Computing* 18: 507–516.

Wang, Y., and X. Ma. 2015. “A General Scalable and Elastic Content-Based Publish/Subscribe Service.” *IEEE Transactions on Parallel and Distributed Systems* 26 (8): 2100–2113.

Wang, Q., C. Wang, J. Li, K. Ren, and W. Lou. 2009. Enabling Public Verifiability and Data Dynamics for Storage Security in Cloud Computing. In *Computer Security–ESORICS 2009*, 355–370. Springer Berlin Heidelberg.

Whitman, R. T., M. B. Park, S. M. Ambrose, and E. G. Hoel. 2014. Spatial Indexing and Analytics on Hadoop. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 73–82. ACM.

Witayangkurn, A., T. Horanont, and R. Shibasaki. 2013. “ The Design of Large Scale Data Management for Spatial Analysis on Mobile Phone Dataset.” *Asian Journal of Geoinformatics* 13 (3). http://a-a-r-s.org/acrs/administrator/components/com_jresearch/files/publications/C4-3.pdf.

Wright, D. J., and S. Wang. 2011. “The Emergence of Spatial Cyberinfrastructure.” *Proceedings of the National Academy of Sciences* 108 (14): 5488–5491.

Wu, K., L. Chen, and Y. Li. 2015. “A Trusted-Based Cloud Computing Virtual Storage System and Key Technologies.” *International Journal of Computers Communications & Control* 10 (4): 579–592.

Wu, H., Z. Li, H. Zhang, C. Yan, and S. Shen. 2011. “Monitoring and Evaluating the Quality of Web Map Service Resources for Optimizing map Composition over the Internet to Support Decision Making.” *Computers & Geosciences* 37 (4): 485–494.

Xia, J., C. Yang, Z. Gui, K. Liu, and Z. Li. 2014. "Optimizing an Index with Spatiotemporal Patterns to Support GEOSS Clearinghouse." *International Journal of Geographical Information Science* 28 (7): 1459–1481.

Xia, J., C. Yang, K. Liu, Z. Li, M. Sun, and M. Yu. 2015a. "Forming a Global Monitoring Mechanism and a Spatiotemporal Performance Model for Geospatial Services." *International Journal of Geographical Information Science* 29 (3): 375–396.

Xia, J., C. Yang, K. Liu, Z. Gui, Z. Li, Q. Huang, and R. Li. 2015b. "Adopting Cloud Computing to Optimize Spatial web Portals for Better Performance to Support Digital Earth and Other Global Geospatial Initiatives." *International Journal of Digital Earth* 8: 451–475.

Xie, J., C. Yang, B. Zhou, and Q. Huang. 2010. "High-performance Computing for the Simulation of Dust Storms." *Computers, Environment and Urban Systems* 34 (4): 278–290.

Xing, F. X., J. Pang, and B. Q. Zhang. 2010. "Application About the Objects Internet Technology in the Modern Agricultural Production." *Agricultural Technology& Equipment* 8: 16–17.

Xing, J., and R. E. Sieber. 2016. "A Land Use/Land Cover Change Geospatial Cyberinfrastructure to Integrate Big Data and Temporal Topology." *International Journal of Geographical Information Science* 30: 573–593.

Xu, Z. 2012. "How Much Power Is Needed for a Billion-Thread High-Throughput Server?" *Frontiers of Computer Science* 6 (4): 339–346.

Xu, J., E. Huang, C. H. Chen, and L. H. Lee. 2015. "Simulation Optimization: A Review and Exploration in the New Era of Cloud Computing and Big Data." *Asia-Pacific Journal of Operational Research* 32: 1550019-1-34.

Yang, C. 2011. Thinking and computing spatiotemporally to enable cloud computing and science discoveries. In 19th *International Conference on Geoinformatics*, June 2011, 1–6.

Yang, C., M. Goodchild, Q. Huang, D. Nebert, R. Raskin, Y. Xu, M. Bambacus, and D. Fay. 2011a. "Spatial Cloud Computing: How Can the Geospatial Sciences Use and Help Shape Cloud Computing?" *International Journal of Digital Earth* 4 (4): 305–329.

Yang, C., Q. Huang, Z. Gui, Z. Li, C. Xu, Y. Jiang, and J. Li. 2013. "Cloud Computing Research for Geosciences." In *Spatial Cloud Computing: A Practical Approach*, edited by C. Yang, Q. Huang, Z. Li, C. Xu, and K. Liu, 295–310. Boca Raton, FL: CRC Press/Taylor & Francis.

Yang, C., C. Liu, X. Zhang, S. Nepal, and J. Chen. 2015a. "A Time Efficient Approach for Detecting Errors in big Sensor Data on Cloud." *IEEE Transactions on Parallel and Distributed Systems* 26 (2): 329–339.

Yang, Y., X. Long, and B. Jiang. 2013. "K-Means Method for Grouping in Hybrid MapReduce Cluster." *Journal of Computers* 8 (10): 2648–2655.

Yang, C., R. Raskin, M. Goodchild, and M. Gahegan. 2010. "Geospatial Cyberinfrastructure: Past, Present and Future." *Computers, Environment and Urban Systems* 34 (4): 264–277.

Yang, C., M. Sun, K. Liu, Q. Huang, Z. Li, Z. Gui, Y. Jiang, et al. 2015b. "Contemporary computing technologies for processing big spatiotemporal data." In *Space-Time Integration in Geography and GIScience*, 327–351. Springer Netherlands.

Yang, C., H. Wu, Q. Huang, Z. Li, and J. Li. 2011b. "Using Spatial Principles to Optimize Distributed Computing for Enabling the Physical Science Discoveries." *Proceedings of the National Academy of Sciences* 108 (14): 5498–5503.

Yang, C., Y. Xu, and D. Nebert. 2013. "Redefining the Possibility of Digital Earth and Geosciences with Spatial Cloud Computing." *International Journal of Digital Earth* 6 (4): 297–312.

Yang, C., X. Zhang, C. Zhong, C. Liu, J. Pei, K. Ramamohanarao, and J. Chen. 2014. "A Spatiotemporal Compression Based Approach for Efficient big Data Processing on Cloud." *Journal of Computer and System Sciences* 80 (8): 1563–1583.

Ye, K., X. Jiang, Y. He, X. Li, H. Yan, and P. Huang. 2012. "vHadoop: A Scalable Hadoop Virtual Cluster Platform for Mapreduce-based Parallel Machine Learning with Performance Consideration." In *2012 IEEE International Conference on Cluster Computing Workshops (CLUSTER WORKSHOPS)*, 152–160.

Yeager, P. S. 2003. "A distributed file system for distributed conferencing system." PhD diss., University of Florida. Accessed November 25, 2015. http://etd.fcla.edu/UF/UFE0001123/yeager_p.pdf.

Yee, K. P., K. Swearingen, K. Li, and M. Hearst. 2003. Faceted Metadata for Image Search and Browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 401–408.

Yue, P., C. Zhang, M. Zhang, X. Zhai, and L. Jiang. 2015. "An SDI Approach for Big Data Analytics: The Case on Sensor Web Event Detection and Geoprocessing Workflow." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8 (10): 4720–4728.

Zelenkauskaite, A., and B. Simões. 2014. Big Data Through Cross-Platform Interest-Based Interactivity. In *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*, 191–196. IEEE.

Zhai, Y., Y. S. Ong, and I. W. Tsang. 2014. "The Emerging "Big Dimensionality"." *IEEE Computational Intelligence Magazine* 9 (3): 14–26.

Zhan, Z. H., X. F. Liu, Y. J. Gong, J. Zhang, H. S. H. Chung, and Y. Li. 2015. "Cloud Computing Resource Scheduling and a Survey of its Evolutionary Approaches." *ACM Computing Surveys (CSUR)* 63: 1–33.

Zhang, F., J. Cao, S. U. Khan, K. Li, and K. Hwang. 2015a. "A Task-Level Adaptive MapReduce Framework for Real-Time Streaming Data in Healthcare Applications." *Future Generation Computer Systems* 43–44: 149–160.

Zhang, Q., Z. Chen, and Y. Leng. 2015. “Distributed Fuzzy C-Means Algorithms for big Sensor Data Based on Cloud Computing.” *International Journal of Sensor Networks* 18 (1–2): 32–39.

Zhang, Q., Z. Chen, and L. T. Yang. 2015. “A Nodes Scheduling Model Based on Markov Chain Prediction for big Streaming Data Analysis.” *International Journal of Communication Systems* 28 (9): 1610–1619.

Zhang, X., X. Li, and J. Chen. 2012. Message from BigDataMR2012, In *2012 Second International Conference on Chairs, Cloud and Green Computing (CGC)*, xxix-xxix, IEEE. doi:10.1109/cgc.2012.136.

Zhang, X., C. Liu, S. Nepal, C. Yang, and J. Chen. 2014. Privacy Preservation Over Big Data in Cloud Systems. In *Security, Privacy and Trust in Cloud Systems*, 239–257. Springer Berlin Heidelberg.

Zhang, F., Q. M. Malluhi, T. Elsayed, S. U. Khan, K. Li, and A. Y. Zomaya. 2015b. “CloudFlow: A Data-Aware Programming Model for Cloud Workflow Applications on Modern HPC Systems.” *Future Generation Computer Systems* 51: 98–110. doi:10.1016/j.future.2014.10.028.

Zhang, M., H. Wang, Y. Lu, T. Li, Y. Guang, C. Liu, E. Edrosa, H. Li, and N. Rishe. 2015c. “TerraFly GeoCloud: An Online Spatial Data Analysis and Visualization System.” *ACM Transactions on Intelligent Systems and Technology (TIST)* 6: 34.

Zhang, L., C. Wu, Z. Li, C. Guo, M. Chen, and F. Lau. 2013. “Moving big Data to the Cloud: An Online Cost-Minimizing Approach.” *IEEE Journal on Selected Areas in Communications* 31 (12): 2710–2721.

Zhang, X., and F. Xu. 2013. Survey of Research on Big Data Storage. In *2013 12th International Symposium on Distributed Computing and Applications to Business, Engineering & Science (DCABES)*, 76–80. IEEE.

Zhao, L., L. Chen, R. Ranjan, K.-K. R. Choo, and J. He. 2015. “Geographical Information System Parallelization for Spatial big Data Processing: A Review.” *Cluster Computing*: 19 (1): 139–152.

Zhao, J., L. Wang, J. Tao, J. Chen, W. Sun, R. Ranjan, J. Kołodziej, A. Streit, and D. Georgakopoulos. 2014. “A Security Framework in G-Hadoop for big Data Computing Across Distributed Cloud Data Centres.” *Journal of Computer and System Sciences* 80 (5): 994–1007.

Zhao, H., S. Yang, Z. Chen, S. Jin, H. Yin, and L. Li. 2012. Mapreduce Model-based Optimization of Range Queries. In *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2487–2492. IEEE.

Zhong, Y., J. Han, T. Zhang, Z. Li, J. Fang, and G. Chen. 2012. Towards Parallel Spatial Query Processing for Big Spatial Data. In *2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW),* 2085–2094. IEEE.

Zhu, T., C. Shu, and H. Yu. 2011. Green Scheduling: A Scheduling Policy for Improving the Energy Efficiency of Fair Scheduler. In *2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)*, 319–326. IEEE.

Zikopoulos, P., and C. Eaton. 2011. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Osborne Media (Chapter 2).