Big Data Meet Green Challenges: Greening Big Data

Article *in* IEEE Systems Journal · September 2016

DOI: 10.1109/JSYST.2016.2550538

4 authors:

Jinsong Wu
University of Chile
100 PUBLICATIONS **262** CITATIONS
SEE PROFILE

Song Guo
The Hong Kong Polytechnic University
300 PUBLICATIONS **1,737** CITATIONS
SEE PROFILE

Jie Li
Duke University
321 PUBLICATIONS **2,264** CITATIONS
SEE PROFILE

Deze Zeng
The University of Aizu
76 PUBLICATIONS **349** CITATIONS
SEE PROFILE

Big Data Meet Green Challenges: Greening Big Data

Jinsong Wu, *Senior Member, IEEE*, Song Guo, *Senior Member, IEEE*,
Jie Li, *Senior Member, IEEE*, and Deze Zeng, *Member, IEEE*

***Abstract*—Nowadays, there are two significant tendencies, how to process the enormous amount of data, big data, and how to deal with the green issues related to sustainability and environmental concerns. An interesting question is whether there are inherent correlations between the two tendencies in general. To answer this question, this paper firstly makes a comprehensive literature survey on how to green big data systems in terms of the whole life cycle of big data processing, and then this paper studies the relevance between big data and green metrics and proposes two new metrics, effective energy efficiency and effective resource efficiency in order to bring new views and potentials of green metrics for the future times of big data.**

***Index Terms*—Big data, data generation, data acquisition, data communications, data storage, data analytics, effective energy efficiency (EEE), effective resource efficiency (ERE), environmental sustainability, energy efficiency (EE), green, green revolution, resource efficiency, sustainability.**

I. Introduction

THE past 20 years, data have been rapidly growing elements of our lives, especially due to the increasing popularity of Internet and relevant applications. More and more data are being produced and analyzed, leading to a new big data era. The characteristics of big data have been discussed for over 15 years since 2011.

1) The first characterization of big data, proposed by D. Laney, META Group, in 2001 [1], includes volume, velocity, and variety, which have been mostly acceptable.
2) The characterization of big data with 5 Vs, i.e., volume, variety, velocity, veracity, and value, was discussed in 2011 [2].
3) The characterization of big data with 6 Vs, i.e., volume, variety, velocity, veracity, visualization, and value, was introduced in 2012 [3].
4) Recently even more 8 Vs, i.e., volume, variety, velocity, value, veracity, variability, viscosity, and virality, have been further proposed in 2014 [4].

Manuscript received February 7, 2016; revised March 14, 2016; accepted March 19, 2016. Date of publication May 19, 2016; date of current version August 23, 2016. This research has been partially supported by Proyecto Iniciacion (FCFM) and start funding (DIE) of Universidad de Chile, Strategic International Collaborative Research Program (SICORP) Japanese (JST) - US (NSF) Joint Research "Big Data and Disaster Research" (BDD), Grant-in-Aid for Scientific Research of Japan Society for Promotion of Science (JSPS), and the NSF of China (Grant No. 61402425, 61502439). *(Corresponding author: Jinsong Wu.)*

J. Wu is with the Department of Electrical Engineering, Universidad de Chile, 1058 Santiago, Chile (e-mail: wujs@ieee.org).

S. Guo is with the School of Computer Science and Engineering, University of Aizu, Aizu-Wakamatsu 965-8580, Japan (e-mail: sguo@u-aizu.ac.jp).

J. Li is with Department of Computer Science, University of Tsukuba, Tsukuba 305-8571, Japan (e-mail: lijie@cs.tsukuba.ac.jp).

D. Zeng is with the School of Computer Science, China University of Geosciences, 430074 Wuhan, China (e-mail: deze@cug.edu.cn).

Digital Object Identifier 10.1109/JSYST.2016.2550538

However, whatever the aforementioned characterizations are, not all types of big data are with all the features for anyone of aforementioned characterizations. Nowadays, there are tremendous demands on the green issues and environmental concerns to information and communication technologies (ICT) [5]. This paper would like to explore the correlation between the trend of big data era and that of green revolution via investigating how to green big data systems.

In this paper, our main contributions are listed as follows.

1) In Section II, we summarize basic features of big data and relevant systems with extensive discussions. In Section III, we clarify the green issues through historical reviews, and show our definition of green concept, and start some initial discussions on the relation between big data and green issues.
2) Section IV presents an introduction of greening big data as the prediscussions to the extensive survey on how to green big data in terms of the whole life cycle of big data processing in Section V–VII.
3) In Section VIII, we study the relevance between big data and green metrics and propose two new metrics, effective energy efficiency (EEE) and effective resource efficiency (ERE), which may bring new viewpoints and potentials of green metrics for the future investigations of big data.
4) Section IX outlines several future promising research challenges to promote the greenness of big data, from data acquisition, storage and processing, respectively.

II. Big Data

Big data refer to not only the realization of information explosion but also technologies, which ensure values generated from the massive of data. Big data have been defined as a broad term for huge or complex datasets that conventional data processing applications are insufficient [6]. In [7], big data were defined in volume as the science in the petabyte era. Further, big data were also defined based on not only high volumes, but also wide varieties and the requirements of high velocity capture, discovery, and analysis [8]. The markets for big data involve three layers: the infrastructure layer (hardware components), the data organization, analytics, and management layer (software components), and the services layer (big data applications). The infrastructure layer, mainly including external storage systems, servers, datacenter networking infrastructure and cloud infrastructure, is the foundation of the big data technologies. The data organization, analytics, and management layer, typically implemented as software, is in charge of storing, processing, and analyzing various structured and unstructured data, which could be in offline, real time, or both. The services layer stands for the big data relevant external interfaces and applications, such as business consulting, project services, integration services, data storage services,

Fig. 1. Three-layer big data.

security supports, and technical trainings. An overview of the three layers for big data is illustrated in Fig. 1.

III. Green Issues

A. Origin and our Definition of Green Issues

This section investigates what are green issues. In many recent literatures, people have addressed green issues simply as the synonyms of either energy efficiency (EE) and/or energy consumption reductions. However, those understandings are incorrect, even if energy relevant issues are recently dominant aspects in the green fields. To understand the relevant concepts, we would explore the relevant history of the term "green," which actually was dated back from the origin of the concept "green revolution."

In a narrow sense, the green revolution refers to a series of research, development, and technology initiatives to increase agriculture (in more narrower sense, rice, wheat, and maize) production, particularly in the developing countries, most notably in the late 1960s. Both the start time of the green revolution, more accurately, the first generation green revolution, have not been uniquely recognized. Many literatures thought that the start time of the green revolution was dated back to the year of 1943 when, in Mexico [9], Rockefeller Foundation started to sponsor a wheat and maize improvement program led by N. Borlaug, called "Father of the Green Revolution," who won the Nobel Peace Prize in 1970 for his significant contributions to support achieving food self-sufficiency and save around a billion people from hunger and famine via highly efficient high-yielding varieties of cereal grains and other agriculture techniques, such as irrigation increases, synthetic fertilizers, and pesticides [10]. Some other literatures thought that the start time of the green revolution was around 1930s due to the pioneer works on high-yielding varieties of the agrarian geneticist N. Strampelli, while some literatures treated N. Strampelli as the forerunner or precursor of the green revolution [11]. We remark that the definitions of second generation green revolution have also not been unique, including nonenergy relevant definitions, such as [12], and energy relevant definitions, such as [13].

There are two different opinions on when and who created the term "Green Revolution." Many people thought based on [14] that the term "Green Revolution" were created in 1968 by the former director of United States Agency for International Development (USAID), W. Gaud, who stated that "These and other developments in the field of agriculture contain the makings of a new revolution. It is not a violent Red Revolution like that of the Soviets, nor is it a White Revolution like that of the Shah of Iran. I call it the Green Revolution." However, some evidences have shown that the term of "Green Revolution" had appeared before 1968. In 1967 [15], W. O. Reichert has recognized the Green Revolution advocated by M. Loomis with the nonviolent and unpolitical feature. In 1962 [16], M. J. Loomis, respected as "the Grandmother of the Counter Culture," addressed the relationship between the impact of R. Borsodi and a number of constructive trends on socio-philosophical thinking, where M. J. Loomis have recognized that R. Borsodi published his opinions about supporting peace via home- and small-scale production on the term of "Green Revolution" [17] on July 28, 1943. In [16], M. J. Loomis explained "Green Revolution" relevant to the "conservation and improved methods of tilling the soil, the use of whole, undevitalized food, good nutrition, and in many cases, simpler diet and simple living."

From the aforementioned discussions, the original meanings of the term "green" are not about energy issues but more relevant to food sustainability as well as the peace of human society opposite to violent wars. In general, green issues have not been uniquely defined, but it is not correct to simply treat green issues as either EE or energy consumption reductions. We would like to state our definition of green issues that green issues refer to those making the world and the components both sustainable and friendly in an environmental, economical, social, and/or technical sense, or in an equation format as

$$\begin{aligned}\text{Green} = {} & \text{Environmentally}/\text{Economically}/\text{Socially}/\text{Technically} \\ & \text{Sustainable} + \text{Environmentally} \\ & /\text{Economically}/\text{Socially}/\text{Technically}\end{aligned}$$

Friendly, which stresses not only sustainability objectives but also friendly characteristics to environments and human societies. We have presented the aforementioned definition of green issues in a number of public presentations since 2011, but it is the first time for us to formally publish this definition. We would have more discussions on the concept of sustainability on [18].

B. Green Issues in the Big Data Era

Big data may require a large scale of data centers with huge computing power and resources. The increase of energy consumption and other resources would result in increased greenhouse emissions and impacts on environments. More specifically, big data generation from sensors, video cameras, and other available data sources greatly stress the existing data generation devices. Big data acquisitions bring a lot of energy consumptions for data collections as well for data transmissions in networks. Next, conventional database software cannot properly handle the storage of various kinds of big data. More storage capacity is designed for big data storage, leading to advanced technology and distributed devices with energy and resource inefficiency. Finally, big data analytics brings the challenges of large-scale data analytics. Parallel or distributed frameworks and architectures are required for big data, which give rise to

Fig. 2. Big data scenario.

more consumptions of energy and other resources. Nowadays, computationally intensive big data analytics and processing may consume a lot of energy and pollute the environment.

Considering the aforementioned concerns, it is really necessary to establish green initiatives for big data life cycle to reduce energy and resource consumptions, to reduce greenhouse gases emission and long-lasting harmful effects on our environments. Fortunately, big data may be huge threatens to not only environmental changes but also enormous opportunities for making environmental improvements. Large scale datasets and analytics increasingly are being used by government agencies, nongovernmental organizations, and private firms to forward environmental protections. For instance, companies like Virginia-based OPower used big data to allow homeowners to measure their consumption against their neighbors. Cities are optimizing the timing of traffic signals to reduce congestions; airports are communicating with planes to increase the efficiency of waiting ground crews; and building managers are using data analysis to cut energy uses by 10–20%.

Improving ecologically oriented efficiency, promoting environmental justices, and tracking climate changes are environmental improvements by the use of big data. Greening big data is an important part of green ICT, which includes green computing [19] and green communications [5], which are two highly overlapped terms. Green computing [19] mainly address investigations and practices of efficient and ecofriendly computing, including designs, productions, uses, and disposals of computers, associated subsystems, and networking devices efficiently and effectively with minimal or without negative impact on the environments. Green communications cover all environmentally relevant issues to communications systems and relevant applications. Product longevity refers to the whole life of a product from the design stage to the end-of-life stage. It is highly expected to have longer product life to reduce necessary product disposals in a certain period, since these may save natural resources already very limited and fast depleted, and more and more product disposals have created more and more wastes, including e-wastes, which may be discarded into land soils and ground waters and include harmful materials such as lead, cadmium, mercury, and chromium.

IV. Introduction for Greening Big Data

As initially mentioned in Section III, there are some relation clues between big data systems and green issues. In 0 Section V–VII, we try to have some extensive investigations in green issues within big data systems. It is hard to process various big data via traditional database management systems. So far, the main IT (information technology) infrastructure for dealing with big data is datacenters. The well-known software frameworks for processing big data are Hadoop MapReduce [20], Google Mapreduce [21], and Amazon Elastic MapReduce [22]. The server clusters running these frameworks in a datacenter will consume a lot of energies. It has been reported [23], [24] that, in a datacenter, the power consumption of severs for data processing is about 45% of energy, and the heating, ventilation, and air conditioning (HVAC) may consume about 30% of energy. With the rapid growth of big data services and applications, datacenter consume more and more energy [25], greening datacenters becomes important and challenge research issue. Although the energy consumption of big data processing is not easily reduced in a datacenter, through the proper arrangements, it is possible to save energy consumption of the HVAC part. The efficient HAVC is helpful in cooling down the datacenters, and improving the big data services.

Actually, besides the high energy and resource consumptions during the data processing stage, a lot of energies and resources are also consumed in both the data generation, acquisition, com-

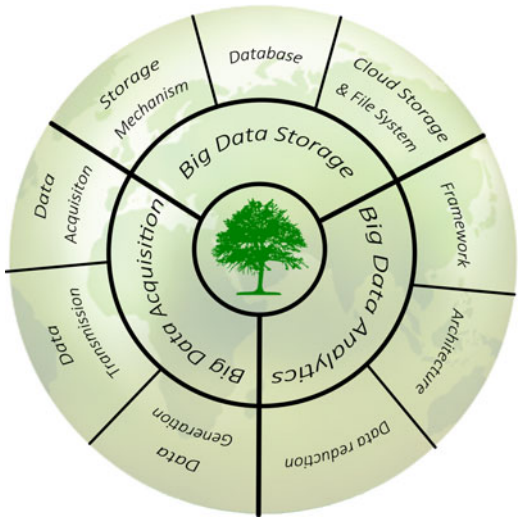


Fig. 3. Greening big data.

munication, and storage stages. As shown in Fig. 2, large volume of data, in size of petabytes or even zettabytes, are continuously generated from a variety of domains, such as industry, agriculture, home appliance, stock market, social network, and so on. Especially, it is inevitable that a large number of Internet of things (IoT) devices will be connected in order to accelerate the convergence of cyber and physical world. Former Cisco CEO J. Chambers predicted that there will be 500 billion connected devices by 2025. These devices may even become dominant data sources in the future big data. Many geodistributed datacenters from large entrepreneurs such as Google, Amazon, and Microsoft have already deployed worldwide. Besides, many companies also established their own private clouds, as complementary to public clouds, to deal with their personal data. Nevertheless, the collected data shall be efficiently acquired, transmitted and stored in these datacenters for further processing or analytics. Each stage in the data life cycle consumes ineligible amount of energy and resource with the consideration of high-volume big data. As a result, significant efforts shall be contributed to the data acquisition, communication, and even storage from the aspects of greenness. This has already raised wide concern in the research community and lots of methods toward greening big data have already been available in the literature. In Section V–VII, we summarize and discuss greening big data on different stages, including big data acquisition, big data storage, and big data analytics, during the whole life cycle. An overview of our discussions in this paper is summarized in Fig. 3.

V. Greening Big Data Generation, Acquisition, and Communications

This section would like to the issues on generation, acquisition, and communications of big data, which all are relevant to a general sustainability issue on how to deal with high volume of big data. One direction on handling high volumes is big data reduction, for which there are two large categories of solutions.

1) *Lossy Reduction*
   a) *Reduce or discard unimportant or useless data:* As a recent progress, recent investigations by Ready Labs, Inc., and Simon Fraser University have showed Adblock Plus, a popular, open source, ad-blocking Internet browser extension, may significantly reduce network data demands where 25.0% reduction in bytes downloaded and 40.0% when video traffic considered in isolation are achieved [26].
   b) *Lossy compression:* This could be supported by many source coding algorithms [27] and other signal processing methods such as compressive sensing.
2) *Lossless Reduction*
   a) *Lossless compression* [27]: Lossless compression typically is used for the individual information or files in an organization.
   b) *Deduplication* [28]: Data deduplication, a specialized compression, performs across all the storage system to discover and remove duplicate data and an index maintains tracking exactly the removal so that the information can be re-accessed when necessary.
   c) *Similarity-based compression methods:* As the relevant theoretical studies to similarity-based compression, [29] defined the compression-based dissimilarity measure (CDM) as

$$\text{CDM}(x,y) = \frac{C(xy)}{C(x)+C(y)} \tag{1}$$

   where x and y stands for two strings, $C(x)$ is defined as the size of the compressed x for a given compression, and $C(xy)$ is the size of the compressed string for y concatenated to x. Further, [30] introduced the concept normalized compression distance (NCD) for a distance metric between two strings, which is defined as

$$\text{NCD}\,(x,y) = \frac{C(xy) - \min\left(C(x), C(y)\right)}{\max\left(C(x), C(y)\right)}. \tag{2}$$

   Both NCD and CDM are based on the concept of Kolmogorov complexity [31]. Recently, a fast compression similarity measure, namely normalized dictionary distance, was proposed based on the idea of dictionary [32]. The fast compression distance (FCD) [33] has been proposed to reduce the high complexity of the previous methods, such as NCD. A new faster similarity measure, called weighted FCD, was most recently proposed in [34].

The aforementioned approaches could be also jointly used. Xia *et al.* [35] discussed a joint approach with both compression and deduplication.

A. Big Data Generations

Data generation is the first step of big data. According to [36], main sources of big data are enterprise data, IoT data, and biomedical data. Not all raw data generated are useful for extracting values, and excessive data generation will cause

huge burden in energy and resource consumptions, thus it is necessary to achieve effective big data generations toward green objectives.

1) Enterprise Data: Recent study [37] has indicated the enterprise data in the world have been the main source of big data. The most known enterprises with huge data are Internet focused companies, such as Facebook, Google, and Tencent, which produce huge data together with huge relevant benefits every day. However, those excessive data generations require huge energy consumptions, which force the enterprises exploiting green initiatives to deal with big data generations. For example, the vast amounts of data being generated could be reused by devices themselves to maximize the EE. To be specific, enterprise servers that run on electricity could generate additional electrical information that could be analyzed to determine the power management mode.In the context of this, servers not only generate massive data but also may save energy based on these data.

2) IoT Data: IoT [38] is an important source of big data. Most IoT data gathered by radio frequency identification (RFID) and sensor network technologies exist in many industries, such as agriculture, traffic, transportation, medical care, public departments, and families, and so on. In sensor networks, IoT devices may be embedded with sensors. From the perspective of green initiatives, an emerging category of edge devices (smart phones) are potentially utilized in data generation, since these devices can reduce the deployment of corresponding sensors, which is considered environmentally friendly. For example, [39] introduced mobile crowdsensing, where individuals with mobile devices collectively share data and extract information to benefit IoT applications. Utilizing mobile devices for sensing may greatly reduce the manufacture and deployment costs. RAMSES [40] considered using existing portable devices for IoT data generation in the hardware design of devices, and through implementing an RF energy-harvesting circuit to wirelessly deliver sensor data to standard RFID readers, the system consumes less energy in the process of data generation.

3) Biomedical Data: Over the past several years, bioinformatics has become an all-inclusive term for everything relevant to both computer and biology sciences, so that biomedical areas may also bring huge data with the features of large volumes, complicated structures, high dimensionality, biological concepts involvements, and insufficient data modeling, which may be also supported by green initiatives in data reductions, scalability improvements, and energy consumption reductions.

B. Big Data Acquisition

With regard to networking in big data acquisition, a lot of works have focused on data aggregation and energy-efficient data acquisition techniques. Besides the general discussions in the start of this section, we have further discussions on this topic. From the perspective of software technologies, there are a lot of approaches in effective data collections. For example, [41] exploited data compression technology, which is capable of shrinking the volume of transmitted data, resulting in fewer energy consumptions. By defining a new transfer time $t_{\text{total}} = t_{\text{total}} = t_{\text{compression}} + t_{\text{transfer}} + t_{\text{decompression}}$, the compression decision can be made via comparing the original transfer time t_{original}. Compression will be applied if and only if $t_{\text{original}} > t_{\text{total}}$. Pioneering researchers have already proposed many different methods to reduce the data volumes so as to save unnecessary energy consumption. These methods can be generally categorized into lossy reduction, lossless reduction, and data aggregation.

1) Lossy Reduction: For data acquisition in WSNs, Liu *et al.* [42] developed a generic framework on how to explore spatio-temporal correlation in wireless sensor network so as to achieve energy-efficient data collections, where the sensor nodes receive their local measures of interest, such as temperature and light intensity, to the sink node continuously. The data sequence obtained at each sensor node creates a time series. Two time series $X\{x_1, x_2, \cdots, x_q\}$ and $Y\{y_1, y_2, \cdots, y_q\}$ are assumed to be trend-t-dissimilarity if

$$\frac{q_1}{q} < t \tag{3}$$

where q is the total number of the pairs (x_i, y_i) for the time series and q_1 is the number of pairs that satisfy $\nabla x_i \times \nabla y_i \geq 0, \nabla x_i = x_i - x_{i-1}, \nabla y_i = y_i - y_{i-1}, 1 \leq i \leq q-1$. According to the dissimilarity between the two time series, a clustering algorithm was proposed to achieve energy-efficient and continuous data collection. To achieve efficient data collection, [43] proposed a principle that limits the unnecessary collection of personal data to greatly reduce the relevant energy consumption. Once the purpose for which data are collected is known, collected data should be just efficient enough for that purpose.

2) Lossless Reduction: Marcelloni *et al.* [44] exploited the high correlation between consecutive samples collected by a sensor and proposed an entropy compression-based data sampling. A recent study [45] propose MinDiff for Gaussian distributed data compression using a dictionary and local prediction method.

3) Data Aggregation: Data aggregation explores in-network processing during the data routing and applies different data operations (e.g., maximum value, minimum value, and average value) to reduce the data volume, and hence, energy consumption. Note that there are some overlaps between the aggregation techniques and data compression techniques but basically they are different. The data aggregation scheme [46] applies compressed sensing for both recovery fidelity and EE in WSNs. The compressed sensing decoding process reconstructs the signal as $\hat{\mu} = \Psi\hat{\omega}$, where μ is an n-dimensional signal and ω is the optimal solution to the convex optimization problem.

$$\min_{w \in R^n} \|\omega\|_{l_1} \left(= \sum_i |w_i| \right) \text{ subject to } \mu = \Phi\Psi\omega. \tag{4}$$

This novel aggregation scheme enables to investigate the minimum-energy compressed data aggregation. In [47], ergelt *et al.* proposed an aggregation strategy to efficiently use the energy and dealing with large data volumes. A recent survey on data compression algorithms for WSNs was reported in [48]. Interesting readers may also refer to [49] for distributed data aggregation algorithms.

C. Big Data Communications

Upon the completion of raw big data generation, data will be collected and transferred through wired or wireless networks for processing or storage. This phase also incurs unprecedented energy consumptions, quality of service issues, and other sustainability issues in big data era. This challenge also has already been widely addressed in the literature.

1) Network Topology: In WSNs, numerous sensors can produce a significant volume of the data but also incur severe energy consumptions. How to optimize network structure as well as data routing for EE also plays an important role in greening big data. During the recent years, many energy-efficient routing protocols have been proposed for WSNs. For example, Takaishi *et al.* [50] utilized the mobility of the sink node to help the data collection for EE.

2) Routing: Based on the maximization of network utility, [51] proposed a fully distributed asynchronous flow control algorithm in WSN.

$$\min \sum_{s \in S} U_s(x_s)$$
$$\text{s.t:} \quad \mathbf{R}^{''}\mathbf{x} \leq \mathbf{c} \tag{5}$$

where $U_s(x_s)$ refers to the utility of flow s under transmission rate x_s, $\mathbf{R}^{''}$ is the generalized routing matrix, and $\mathbf{c}$ is the link capacity matrix. Static routing cannot efficiently balance traffic for big-data applications, [52] tried to reduce each flow bandwidth via a distributed-adaptive-routing algorithm with minimal out-of-order packet delivery. Pantazis *et al.* [53] recently conducted a survey on energy-efficient routing protocols in WSNs. In [54], Bergler *et al.* presented an approach to reduce the energy costs for datacenters through selecting routing with the current cheapest energy costs. To more effectively and efficiently utilize multiple cores in routing the big data transfers, [55] proposed to parallelize data transfers via using each core in the routers to calculate a separate shortest path.

3) Network Infrastructure: Besides efforts on existing IP networks, researchers also have investigated network infrastructure revolution for big data. Recent advances of software defined networking and optical switching technology make it possible to program the network stack all the way from physical topology to flow level traffic control. In [56], the EE of the networking components in cloud-based environments was provided. Ricciardi *et al.* [57] provided the modeling and the cross-layer optimization approach for the EE and energy-awareness Internet design. Wang *et al.* [58] combined a software-defined networks (SDNs) controller with optical switching to tightly integrate application and network control, which may greatly enhance application performance with small configuration overheads. Another work addressing SDN [59] tried to realize dynamic and highly efficient bulk data transfer via a geodistributed datacenter system and an SDN architecture, where data transfer demands were modeled as delay tolerant migration requests with different deadlines. To reach different levels of optimality and scalability in an online fashion, the authors discussed three algorithms for optimal schedules. In [60], the authors utilized the approach of information-centric networking (ICN), where name-based data retrieval and in-network caching are applied. In existing ICN, content centric network (CCN) is not able to efficiently utilize the caches for data sharing due to the use of on-path caching, and network of information (NetInf) shows the resolution latency to retrieve data. Li *et al.* [60] integrated the strengths of CCN and NetInf with the use of information islands and management plane for direct data retrieval and global data discovery, respectively, to facilitate big data sharing. An aggregatable name-based routing (ANBR) was proposed to naturally allow consumers to search the closest copy of information [60].

4) Scheduling: In [61], Ren *et al.* proposed a better communication scheduling strategy CLSA for smartphone downloading to achieve low energy consumption with tolerances of some transmission delays. Wang *et al.* [62] studied the multiple bulk data transfers scheduling problem for network congestion reduction, for which the authors proposed to lexicographically minimize the congestion among all links in datacenters.

VI. Greening Big Data Storage

This section discusses greening datacenters, cloud computing, and other data storage issues. A data center is a computing resource in which a facility is used to house computer systems and associated components. The architecture of datacenter can indirectly influence the consumption of energy. A good datacenter design requires less amount of energy consumption. According to the United States Department of Energy, energy-efficient datacenter designs may impact many technical areas, such as ICT systems, environmental conditions, air management, cooling systems, and electrical systems. Big data storage systems requires the consumptions of both energy and resources, thus green storage should have less energy consumptions, resource consumptions, or both. We now consider the following aspects for green big data storage.

A. Cloud Storage and File System

Cloud computing is a recent popular and powerful paradigm providing various collections of virtualized resources as services suitable for big data storage and processing. As big data applications proliferate, distributed file systems store the data in distributed large-scale nodes and devices. Although cloud computing may help achieve higher energy and resource efficiencies, a big challenge for cloud computing infrastructures is to operate together with the increasing costs for energies and resources. Nowadays, EE improvements are important targets toward green objectives for distributed file systems to store and manage big data. For example, Kaushik and Bhandarkar [63] proposed an energy-conserving logical multizoned variant of Hadoop distributed file system (HDFS) to manage data processing in Hadoop cluster. Idleness of servers may be enabled via data classification, which allowed to use aggressive inactive power modes in central processing unit, disks, and dynamic random access memory (DRAM). In addition, Lightning [64] is a self-adaptive Commodity Green Cloud Storage that dynamically configures the servers into hot and cold zones, where, similar to GreenHDFS [63], the servers in cold zone are transitioned to inactive power modes and the relevant energy saving

significantly save the operating costs of the datacenter. Recently, Bostoen *et al.* [65] made a survey for the research efforts on power-aware enterprise storage systems over a decade. Negru *et al.* [66] reviewed the methods and technologies and presented some of the key research challenges for energy-aware datacenter operations for cloud storage. The authors mentioned that the profitability may not always be expressed with respect to revenue but in terms of bio and environment friendliness.

B. Database

To handle big data, large databases should be partitioned across different servers or nodes, which may overall consume huge energy consumptions. Mardamutu *et al.* [67] provided a comprehensive critical analysis on EE issues for green databases and proposed a redundancy-based solution. According to [68], the first step to achieve green databases is to understand the energy requirements of data warehouse to develop a relevant plan for EE. Then, it is necessary to optimize the server rooms and exploit virtualization to increase performance. Finally, authorities get the feedback via analyzing the developed system. Chaudhary *et al.* [68] studied the tradeoffs between the performance and the energy consumption characteristics of analytical queries based on various database cluster designs. Based on empirical experiment results, the authors proposed a model considering key bottlenecks of EE in a parallel database management system. Wu *et al.* [69] jointly considered hardware accelerator for range partitioning (HARP) and a streaming framework with a seamless execution environment for streaming accelerators, such as HARP, to provide an order of magnitude improvement in partitioning performance and energy for high-throughput, energy-efficient data partitioning and processing, which is critical in manipulating large datasets.

C. Storage Mechanism

In addition to green storage systems, there have been lot of efforts on green storage mechanisms.

One direction is on intelligent storage resource management. A particular appeal is on the power-saving from the perspective of disks. However, it has been reported that, in a datacenter, the average idle period for a server disk is very small compared with the time for spinning down and up, which significantly constrains the performance of disk power management schemes [70]. Zhu *et al.* [70] proposed several power-aware storage cache management algorithms to improve energy savings for the disk power management schemes. Pinheiro *et al.* [71] introduced diverted accesses to leverage the redundancy in storage systems to save disk energy. Another effort [72] tried to enable the storage system to turn OFF a large fraction of disks without unacceptable performance degradation. For real-time and data-intensive applications, Liu *et al.* [73] proposed a novel distributed energy-efficient scheduler, consisting of three main components, i.e., energy-aware ranking, performance-aware scheduling, and energy-aware dispatching, to achieve energy savings via seamlessly involving the process of scheduling tasks in data placement.

Another direction is on data column reduction to save storage resources for energy saving. Lossless compression [27] and deduplication [28] are two kinds of lossless reduction. Compression typically is used for the individual files in a company. While, data deduplication, a specialized compression, execute across all the storage system to identify and remove duplicate data and an index maintains tracking exactly what has been removed so that the information can be put back when a user needs to access a file. Xia *et al.* [35] also discussed a joint approach with both compression and deduplication.

VII. Greening Big Data Analytics

Big data analytics may analyze huge volumes of data that conventional analytics and business intelligence solutions cannot handle. The process of data analytics are normally accompanied with lot of computing workloads, which may be time consuming and energy and resource inefficient. There have been many efforts addressing green issues in big data computing and processing. We now consider how to construct green big data analytics from green big data processing architecture and data analytics framework.

A. Green Big Data Processing Architecture

1) Innovative Architecture: Typically, big data are delivered to cloud for analysis, thus the EE of datacenter architectures should be considered. Unlike high performance computing systems, which generally focus on maximizing raw computing power, big data clusters are generally designed to maximize the EE to manage large datasets [74]. Gu *et al.* [75] investigated and compared the energy consumption and the execution time for a typical Hadoop-based big data application executed in a traditional Xeon-based cluster and an Atom-based (microserver) cluster. Liu *et al.* [76] presented the GreenCloud architecture to reduce datacenter power consumption without sacrificing the performance for users, which enables comprehensive online-monitoring, live virtual machine (VM) migration, and VM placement optimization. Convey [77], [78] was designed to excel at tackling the data-level parallelism with power efficiency. The new Convey architecture can exploit massive degrees of parallelism to enhance development productivity. Huawei [79] proposed a high throughput computing datacenter architecture with PB-level data processing capacity and high EE. In [80], Dben *et al.* provided a preliminary investigation on obtaining energy saving via injecting inexactness or approximation into the hardware architecture of a computing system.

Big data processing naturally requires high parallelism for fast processing. In [81], Kanoun et al. proposed a novel low-power many-core architecture to support the dynamic data-driven nature of stream mining applications while limited power constraints. In [82], a program optimization issue was studied to handle the challenges raised by big data, and the authors discussed using multi/many cores, wide single instruction multiple data and the role of dynamic optimization in the context of big data and cloud-based architectures. The inherent parallel ability of graphics processing unit (GPU) for big data also attracted much attention in the community. In [83], Wang *et al.* proposed

the energy-efficient implementations of GPUs in both the training phase and the operation phase of neural networks for big data analytics. In [84], Park *et al.* presented a special function unit, called memory fast-forward (MFF) unit, to reduce a large number of memory requests for big data workloads in GPU. The two key functions of memory pointer chasing and memory request coalescing for the proposed MFF unit contribute to reducing memory stall time as well as enhancing the real utilization of memory bandwidth via removing duplicate memory traffics, thereby improving performance and EE.

2) Server Management: To lower the environmental impact and power consumption of datacenters, the authors in [85] proposed to comanage both supply and demand via profiling datacenter power consumption, which may be applicable for an enterprise datacenter. In [86], the authors reported an integrated approach for the implementation of a Environment Compatible datacenter, ECCO, which is able to automatically and dynamically redefine the set of active resources in order to drastically lessen the energy consumptions without sacrificing the needs of users. In [87], Schroder and Nebel presented an interesting load and power management method for the server allocation in datacenters. Mastelic *et al.* [88] provided an overview of the infrastructure of the cloud computing paradigm to improve EE. In [89], Lin *et al.* investigated how much can be saved through dynamically scheduling the datacenter for big data cloud services by turning OFF servers during some periods via an online algorithm.

3) System Model: In [90], a brief overview was provided on the benchmarks for datacenter EE, power and energy measurements, benchmarking, and analysis in big data processing. In [91], a brief review of measurements and performance evaluation in term of EE of datacenter for big data processing was provided. The energy effectiveness metric was given by [91]

$$\text{Energy Effectiveness} = \frac{\text{IT Equipment Power Consumption}}{\text{HVAC}}.$$

As HAVC, e.g., cooling, plays a critical role in energy effectiveness, many novel cooling methods have already been proposed. In [92], Wei and Ren introduced their green cloud computing proposal, environmentally opportunistic computing, where information technology resources were working together with existing facilities consuming dominate energy to dynamically adjust process throughput, thermal movement, and obtainable cooling. In [93], Ren intended to improve the datacenter water efficiency. Since datacenter water efficiency varies by location and also over time, the authors proposed a geographical load balancing (GLB) algorithm, called GLB for water sustainability (GLB-WS), which dynamically schedules workloads to improve the overall water usage effectiveness (an emerging metric for quantifying datacenter water efficiency) without violating the electricity cost constraint. Berral *et al.* [94] presented an energy-aware scheduling framework for datacenter using machine learning for big data processing. In [95], a green cloud computing modeling method was provided to build system software environments. In [96], Pelley proposed an analytic framework to model the total power consumption of datacenter. In [97], a set of formal models were provided to estimate sustainability impact and dependability metrics in datacenter infrastructures based on an integrated environment.

4) Renewable Energy: Besides optimization on brown energy, green energy, a.k.a., renewable energy, also attracted much attention. Addis *et al.* [98] discussed the energy-aware joint management of brown and green energy for networks and cloud infrastructures. In [99], Liu *et al.* tried to achieve the reduction of electricity cost and environmental impact via a holistic approach to jointly consider renewable supply, dynamic pricing, and cooling supply with IT workload planning to enhance the overall sustainability of datacenter operations. In [100], Li *et al.* presented a power provisioning scheme called Oasis to scale power-/carbon- constrained datacenter servers to in a economical and green way.

B. Green Big Data Processing Frameworks

It is important to rapidly extract critical information from massive data so as to bring values for enterprises and individuals. Different types of frameworks could run different types of analytics. For example, MapReduce [101]-based framework like Hadoop [102] was aimed for batch-oriented processing. Storm framework [103], [104] was used to hand stream processing. Drill [105] plays a role in interactive ad hoc query and analytics. From the perspective of green computing, it is necessary to build efficient framework for big data analytics. For example, [106] established green MapReduce to reduce energy consumption while maintaining a low task response time. Goiri *et al.* [107] proposed a MapReduce framework, GreenHadoop, for a datacenter energized by a solar array and the electrical grid, which could predict the amount of solar energy to continue and schedule MapReduce jobs to maximize the use of green energy. In [108], Cavallaro *et al.* studied the various smart data analytics methods that take advantage of the support vector machines machine learning algorithm and parallelization approaches to solve the big data processing problems. In [109], a framework with a greedy algorithm, energy-aware MapReduce scheduling algorithm, was proposed to improve the EE of MapReduce applications while remaining the service level agreement. To lower cooling energy consumption and ensure thermal reliability of the servers, [110] introduced a new efficient data-centric approach T*, which makes proactive and thermal-aware file placement to achieve thermal-aware job placement in the big data analytics compute model with the knowledge of the uneven thermal-profile and differences between the thermal-reliability-driven load thresholds of different servers, and the differences in arrival rates of computational jobs, sizes, and evolution life spans of the big data in the cluster. In [111], a Green cloud enabled framework was proposed as the energy-efficient way of a minimal discharge and rectification of the problem of high carbon production so as to increase the profit margin.

C. Data Reductions for Big Data Computing and Processing

One of important challenges to the attainability of big data processing is the high volume data, especially more typically with high dimensions. There have been a number of works toward processing load reductions. Note that we have discussed

TABLE I
DATA REDUCTION TECHNIQUES FOR GREENING BIG DATA

Work	Application Stage	Lossless	Reduction Ratio	Computation Complexity	Single/Multiple Data	Main Concept
[26]	Communication	No	25%	n/a	single	Adblock Plus based reduction
[42]	Acquisition	No	42%–97.3%	$O(n^2 \log n)$	single	Exploring spatio-temporal correlation
[44]	Acquisition	Yes	66.99%–67.33%	42% of S-LZW	single	Entropy compression
[45]	Acquisition	Yes	20%–80%	20%-30% larger than HC	single	Dictionary and local prediction
[46]	Acquisition	No	50%–75%	$O(n^3)$	single	Aggregation and compressed sensing
[47]	Acquisition	No	$\sim$ 30%	n/a	single	Database-orientation
[35]	Storage	Yes	$\sim$80%	1/4 of super-feature approach	single	Deduplication and delta compression
[118]	Analytics	Yes	1000x	n/a	single	FPGA-acceleration
[121]	Acquisition	Yes	$\sim$ 45%–75%	12 instructions for a saved bit	multiple	Lossless entropy compression

some data reduction issues in Section V. Here, the further relevant issues still exist in the stage of big data computing and processing. In [112], Lathauwera and Vandewalleb, reviewed a multilinear generalization of the singular value decomposition (SVD) and the best rank-(R_1, R_2, R_N) approximation for higher order tensors, which may efficiently support dimensionality reduction in higher order signal processing. Based on multilinear algebra, Wang and Ahuja [113] proposed a tensor rank-one decomposition via decomposing multidimensional data into a collection of rank-1 tensors, which was used in image sequence compression of higher quality images to obtain the same compression ratio as principal component analysis, while authors [114] also proposed rank-R tensor approximation for the same objective. Symeonidis *et al.* [115] proposed a unified framework for modeling the three kinds of entities, users, items and tags, in social tagging systems as a three-order tensor, where latent semantic analysis and dimensionality reduction was realized based on higher order singular value decomposition.

In 2009 [116], National Science Foundation, USA, organized a workshop to summarized the relevant issues and challenges of tensor-based computation and modeling, which stressed the importance of tensor-based mathematics, as important parts of multilinear algebra, on the solving the problems of multidimensional datasets. Lua *et al.* [117] reviewed multilinear subspace learning for dimensionality reduction of multidimensional data directly based on their tensorial representations. Jun *et al.* [118] proposed ZIP-IO, an efficient framework for field-programmable gate array (FPGA)-accelerated compression. To interactively visualize big data directly for large-scale computing, Bi *et al.* [119] proposed a parallel compression method, proper orthogonal decomposition, to reduce the data size with low computational cost. Wang *et al.* [120] utilized a wavelet transform to decompose and approximate remote sensing big data with the large scale in the space domain, the correlation in the spectral domain, and the continuity in the time domain via a two-component Gaussian mixture model to check whether the density function of wavelet coefficients for a big dataset is with peaks at zero a heavy tailed shape.

As data reduction can be applied across different stages for greening big data, we summarize some related studies with respect to their characteristics (i.e., application stage, lossless or not, main concepts) in Table I.

VIII. DISCUSSIONS ON THE RELEVANCE AMONG GREEN MEASURES AND BIG DATA

When discussing green issues, many people possibly first think about the popular metric, EE. An important question is what is the proper definition, meaning, and role of EE in the times of big data. Actually, the disputes on the definitions of EE have existed for a long time, and up to now, no consistent agreements are available. In 1996, M. G. Patterson provided a comprehensive summary on the definitions of EE [122], which we consider general enough even at present. The general definition of EE could be [122]

$$\text{EE} = \frac{\text{Useful output of a process}}{\text{Energy input into a process}}. \tag{6}$$

Remark that sometimes the EE is defined via exchanging the positions of the denominator and nominator in (6) as

$$\text{EE} = \frac{\text{Energy input into a process}}{\text{Useful output of a process}}. \tag{7}$$

The author further summarized the definitions of EE into four main groups [122].

1) *Thermodynamic:* The outputs and inputs of (6) in this case are completely based on measurements derived from the science of thermodynamics.
2) *Physical Thermodynamic:* In this case, the inputs of (6) are thermodynamic units, while the outputs of (6) are physical units.
3) *Economic Thermodynamic:* In this case, the inputs of (6) are thermodynamic units, while the service delivery outputs of (6) measured in market prices.
4) *Economic:* In this case, both the energy inputs and service delivery outputs are purely in terms of market monetary values.

In the recent works of ICT, physical-thermodynamic definitions of EE for the time t are typically used as [5]

$$\eta(t) = \frac{N_B(t)}{E(t)} \tag{8}$$

where t refers to a specific time or time slot, $\eta(t)$ is the EE at t, $N_B(t)$ is the number of bits of the information sequences, and $E(t)$ is the quantity of the relevant energy. Besides the general definition of EE for ICT, there have been also quite a few specific definitions of EE for ICT in different scenarios [123]–[126]. All

those existing definitions of EE have not specifically considered big data issues.

As discussed in [122], for the fairness of comparisons, energy quality issues have to be considered using a quality factor β, since different types of energy sources are with different qualities. Further, in the context of big data, the degrees of importance for different data are different at t, which are expressed as an important factor α. When the scenarios are communications and networking, the degrees of importance may be expressed as quality of service parameters. Then, we propose to rewrite the definition of EE (8) as

$$\eta_e(t) = \frac{\alpha N_B(t)}{\beta E(t)} \tag{9}$$

where we call $\eta_e(t)$ effective energy efficiency (EEE) to distinguish from the convectional definitions of EE. If we compare two different EEE $\eta_e^{(1)}(t)$ and $\eta_e^{(2)}(t)$ for two data sequences, which are the same, $N_B^{(1)}(t) = N_B^{(2)}(t), E^{(1)}(t) = E^{(2)}(t), \beta^{(1)} = \beta^{(2)}$, except $\alpha^{(1)} > \alpha^{(2)}$, it is clear that $\eta_e^{(1)}(t) = \frac{\alpha^{(1)} N_B^{(1)}(t)}{\beta^{(1)} E^{(1)}(t)} > \eta_e^{(2)}(t) = \frac{\alpha^{(2)} N_B^{(2)}(t)}{\beta^{(2)} E^{(2)}(t)}$. We assume that $\alpha^{(1)} > \alpha^{(2)}$ shows that the sequence (1) is more important than the sequence (2) at t. For fairness, we would like to have $\eta_e^{(1)}(t) = \eta_e^{(2)}(t)$ for $\beta^{(1)} = \beta^{(2)}$, which could be achieved if $\frac{\alpha^{(1)} N_B^{(1)}(t)}{\alpha^{(2)} E^{(1)}(t)} = \frac{N_B^{(2)}(t)}{E^{(2)}(t)}$, and in this case, $\frac{N_B^{(1)}(t)}{E^{(1)}(t)} < \frac{N_B^{(2)}(t)}{E^{(2)}(t)}$. Thus, $\frac{E^{(1)}(t)}{N_B^{(1)}(t)} > \frac{E^{(2)}(t)}{N_B^{(2)}(t)}$ says that the energy per bit for the sequence (1) is more than that for the sequence (2), in other words, the more important bits should consume more energies at t. This result may inspire many new ways to implement future big data relevant systems. One may ask how to explain this result if all bits of both important and less important bits are actually processed using the same EE λ. Let us assume that the number of bits for the data sequence 2 is partitioned into two parts, $N_{B(1)}^{(2)}(t)$ and $N_{B(2)}^{(2)}(t)$ where $N_B^{(2)}(t) = N_{B(1)}^{(2)}(t) + N_{B(2)}^{(2)}(t)$, and further assume that

1) the actual energy per bit for the $N_{B(1)}^{(2)}(t)$ bits is λ and the actual energy per bit for the $N_{B(2)}^{(2)}(t)$ bits is 0, which enables $E^{(2)}(t) = \lambda N_{B(1)}^{(2)}(t) + 0 N_{B(2)}^{(2)}(t) = \lambda N_{B(1)}^{(2)}(t)$ and $\frac{E^{(2)}(t)}{N_B^{(2)}(t)} = \lambda \frac{N_{B(1)}^{(2)}(t)}{N_B^{(2)}(t)} < \lambda$. Since not all the bits actually consume energy, we may call $\frac{E^{(2)}(t)}{N_B^{(2)}(t)}$ as virtual EE. This case of the $N_{B(2)}^{(2)}(t)$ bits may be implemented using delayed transmissions for the scenarios of communications and networking or power shutdown for the scenarios of data storage;
2) the actual energy per bit for all the $N_B^{(1)}(t)$ bits is λ, which enables $E^{(1)}(t) = \lambda N_B^{(1)}(t)$ and $\frac{E^{(2)}(t)}{N_B^{(2)}(t)} = \lambda \frac{N_{B(1)}^{(2)}(t)}{N_B^{(2)}(t)} < \lambda$.

Using the similar strategy, comparing with the concept of resource efficiency

$$\lambda(t) = \frac{N_B(t)}{R(t)} \tag{10}$$

where $N_B(t)$ is the number of bits of the information sequence, and $R(t)$ is the quantity of the relevant resources, we also propose the concept of ERE as

$$\lambda_e(t) = \frac{\alpha N_B(t)}{\theta R(t)} \tag{11}$$

where $\lambda_e(t)$ is the ERE, θ is the quality factor of the resources, and α is the importance factor of the information sequence. Similarly, we could conclude that the more important bits should occupy more resource at t. Based on the similar strategies, the aforementioned new proposals of EE for big data era could be easily extended for other different definition formats of EE according to different scenarios [123]–[126].

As a remark, we would like to stress the importance of resource efficiency or more general ecological efficiency. Remark that the term resources here typically does not refer to energy but more general meanings, such as spaces, materials, devices, and so on. In literature of recent years, although EE and management have been highly investigated [5], resource efficiency has highly neglected or much less studied. In 1988 [127] and 1996 [128], H.T. Odum created the concept of energy accounting. In 1992 [129], William E. Rees proposed the concept of ecological footprints. Although energy accounting and ecological footprints are different approaches, both of them tried to study substantiality issues via addressing both energy and resources throughput to evaluate the difference between demands of humanity and natural sources. Although ecological footprint methods recently received some criticisms, such as [130], on the application limitations, the authors also agreed to the importance of resource efficiency relevant analysis. The concept of ecological footprint has received significant attentions, which actually criticized the limitations of the popular concept of carbon footprints.

Further, we remark that the concerns on energy, resources, and sustainability have actually been raised by the relevant information with the supports of big data sources, which show the strong correlation among green and big data issues.

IX. Challenge and Open Issues

Although a number of efforts have been relevant to greening big data, many issues are still under-investigated. In this section, we outline several promising future research directions on data acquisition, storage, and processing, respectively.

A. Data Acquisition and Communications

Explore in-network processing to reduce unnecessary data collection A special challenge is on the decision, which essentially is context aware. Consider a scenario where future autonomous cars are capable of sensing their surroundings and cooperate with their neighbors. Definitely there will be vast volumes of data generated by the vehicles at a certain area, but there will be a large information overlapping between the collected data. While, it is nontrivial to detect the duplicated or redundant data in a specific area. Far beyond that, a variety of forms of data may exhibit similar features, and therefore, significant efforts shall be devoted to digging out such phenomena to reduce the collected data volume for energy and resource efficient data

acquisition. Many different compression methods are required to explore spatial-temporal correlation between different data flows for the collected data volume reduction.

Some innovative networking technologies have been proposed recently to welcome big data era. Two representative techniques are known as software-defined networking (SDN) and network function virtualization (NFV). Both intends to promote the flexibility, customability, and managability of networks. SDN and NFV provide a promising direction for energy-efficient data transmission via enabling flexible network resource optimization and allocation. However, at the initial storage of SDN and NFV development, many issues toward green networking are still underaddressed. One one hand, we shall first consider how to use SDN and NFV to enable efficient network management. For example, we shall investigate how to integrate SDN and NFV into existing network infrastructure to promote the energy and resource efficiency of the hybrid networks. On the other hand, how to lower the energy and resource consumption introduced by SDN and NFV themselves shall be also considered. For example, we shall consider how to deploy and migrate network functions adaptively to reduce the energy and resource consumption.

By now, specifically for greening big data, the new designs of communication mechanisms and approaches are still very limited, and many issues in physical layer, media access control layer, network layer, and other upper layers for both wired and wireless networks are to be solved to welcome big data era, such as signal processing methods, control plane issues, network structures, protocols, hardware and software designs, and so on.

B. Storage

It is highly expected to find innovative storage infrastructures that explore a variety of heterogeneous storage resources to reduce the resource and energy consumption in datacenters. For example, some recent studies proposed to exploit vehicles in airport parking lots to provide storage services. However, such paradigms do not support persistent data storage service as the storage resources attached on vehicles are volatile. With the consideration of diverse big data storage requirements, it is beneficial to investigate how to orchestrate the diversity of storage devices towards the objective of greenness.

Recent developments in storage techniques enable superhigh-speed storage devices for large-scale datacenter applications. For example, RAMCloud that keeps all data in DRAM at all times is widely recognized as one feasible way for fast big data processing. However, it is still promising to find out an efficient way to use traditional storage resources like hard disks and the volatile DRAM resources to balance between performance and efficiency, with joint considerations of the inherent characteristics (such as reliability, read/write speed, capacity, power consumption, and so on) of various storage devices.

Another noteworthy thing is on the big data garbage. To pursue the vision of big data, people may aggressively acquire and store various kinds of generated data. However, actually not all these data need to be stored on expensive storage. We shall consider how to eliminate such useless data to avoid unnecessary waste of storage resource as well as the incurred energy consumption. On the other hand, as we already discussed in this survey, the data exhibit both spatial and temporal correlation, which can be explored to eliminate the data redundancy or compress the data volume. However, it is challenging to determine the extent data reduction without incurring negative side effects to data analytics.

C. Data Processing

Similar to storage issues, innovative infrastructures are also demanded for efficient data processing. Mobile cloud communities think that mobile devices shall not be only treated as service consumers but also shall be service providers. The first problem raised is how to orchestrate the infrastructure-based cloud resources and infrastructure-less mobile resources for quality-of-service guaranteed service provision.

With the consideration on high volumes of big data, it is widely believed that the data storage and processing shall be jointly considered for various optimization objectives, such as cost minimization [131]. While, for energy and resource efficiency, we may further take another two important issues, i.e., data acquisition and datacenter cooling. These four aspects are highly correlated with each other and impose deep impact on the performance, resource, and EE. The tradeoff shall be well balanced for efficient and quality of service (QoS) guaranteed big data processing.

As datacenter is the hotspot of big data processing, we shall also seek new electricity provision method for EE. As discussed in this paper, smart grid integrated with renewable power supply, also provides a good way for greening big data. A recent emerging concept, Energy Internet (EI), with the fusion of information technologies and energy technology, is considered as smart grid 2.0 and can better solve the energy crisis and environmental pressure [132]. It is interesting to incorporate EI into big data and investigate how they interact with each other to promote greening big data processing.

X. Conclusion

To discuss the green issues of big data systems, our discussions have addressed relevant green issues throughout the three phases of the life cycle for big data: 1) data generation, data acquisition, and data communications; 2) data storage; and 3) data analysis and processing. Then, we have analyzed the relevances among green measures and big data, discovered that the necessity to adjust the definitions of green measures to respond to the times of big data, and proposed the concepts of EEE and effective resource efficiency.

References

[1] D. Laney, "3D data management: Controlling data volume, velocity, and variety," Application Delivery Strategies, META Group, Feb. 2001.
[2] Dave Beulke and Associates. (2011, Nov.). Big data impacts data management: The 5 Vs of big data. [Online]. Available: http://davebeulke.com/big-data-impacts-data-management-the-five-vs-of-big-data/.
[3] Enterprise Strategy Group. (2012, Aug.). The 6 vs: The BI/analytics game changes so microsoft changes excel. [Online]. Available: http://www. esg-global. com/blogs/the-6-vs-the-bianalytics-game-changes-so-microsoft-changes-excel/.

[4] W. Vorhies. (2014, Oct.). How many V's in big data? The characteristics that define big data. Data Science Central. [Online]. Available: http://www. datasciencecentral. com/profiles/blogs/how-many-v-s-in-big-data-the-characteristics-that-define- big-data.

[5] J. Wu, S. Rangan, and H. Zhang, *Green Communications: Theoretical Fundamentals, Algorithms, and Applications*. Boca Raton, FL, USA: CRC Press, Sep. 2012.

[6] Wikipedia. (2015, Jul.). Big data. [Online]. Available: https://en.wikipedia.org/wiki/Big_data

[7] "Big data: Science in the petabyte era," *Nature*, vol. 455, no. 7209, Sep. 2008.

[8] Info-communications Development Authority of Singapore. (2011). Infocomm technology roadmap. Available: http://www.ida.gov.sg/Tech-Scene-News/Technology/Technology-Roadmap.

[9] J. Cotter, *Troubled Harvest: Agronomy and Revolution in Mexico, 1880-2002 (Contributions in Latin American Studies)*. New York, NY, USA: Praeger, Sep. 2003.

[10] N. E. Borlaug, "The green revolution revisited and the road ahead," Special 30th Anniversary Lecture at The Norwegian Nobel Institute, Oslo for 1970 Nobel Peace Prize Laureate, Norwegian Nobel Inst., Oslo, Norway, Sep. 2000.

[11] S. Salvi, O. Porfiria, and S. Ceccarellia, "Nazareno strampelli, the prophet of the green revolution," *J. Agricultural Sci.*, vol. 151, no. 1, pp. 1–5, Feb. 2013.

[12] P. L. Pingali, "Green revolution: Impacts, limits, and the path ahead," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 31, pp. 12 302–12 308, Jun. 2012.

[13] N. Kulatilaka, "Green revolution 2. 0: A sustainable energy path," *Sustainable Development Insights*, No. 006, pp. 1–8, Oct. 2010.

[14] W. S. Gaud, "The green revolution: Accomplishments and apprehensions," Talk for the Society for International Development, Mar. 8, 1968.

[15] W. O. Reichert, "Toward a new understanding of anarchism," *Western Political Quart.*, vol. 20, no. 4, pp. 856–865, Dec. 1967.

[16] M. J. Loomis, "Children and ourselves: The green revolution," *MANAS J.*, vol. XV, no. 14, pp. 9–11, Apr. 1962.

[17] R. Borsodi, "The green revolution," *Christian Century*, Jul. 1943.

[18] J. Wu, S. Guo, J. Li, and D. Zeng, "Big data meet green challenges: Big data towards green applications," *IEEE Syst. J.*, Mar. 2016.

[19] S. Murugesan, "Harnessing green it: Principles and practices," *IEEE IT Prof.*, vol. 10, no. 1, pp. 24–33, Jan./Feb. 2008.

[20] X. Su and G. Swart, "Oracle in-database Hadoop: when Mapreduce meets RDBMS," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, May 2012, pp. 779–790.

[21] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," presented at the 6th Symp. Operating System Design Implementation, San Francisco, CA, USA, Dec. 2004.

[22] Amazon EMR. (Jul. 2015). [Online]. Available: http://aws.amazon.com/elasticmapreduce/

[23] M. Lyengar and R. Schmidt, "Energy consumption of information technology data centers," Electronics Cooling. (Jul. 2015). Available at http://www.electronics-cooling.com/2010/12/energy-consumption-of-information-technology-data-centers/

[24] K. DeWitt. Tech companies get creative in keeping data center cool. (Aug. 2015). [Online]. Available: http://blog.opower.com/tag/data-centers/

[25] Natural Resources Defense Council. (2014, Aug.). Americas data centers are wasting huge amounts of energy—Critical action needed to save billions of dollars and kilowatts. Issue Brief, IB:14-08-a. [Online]. Available: http://www.nrdc.org/energy/files/data-center-efficiency-assessment-IB.p df.

[26] A. Parmar, C. Dedegikas, M. Toms, and C. Dickert. (2015). Adblock plus efficacy stud. Technical Report, Simon Fraser University. [Online]. Available: http://www. sfu. ca/content/dam/sfu/snfchs/pdfs/Adblock. Plus. Study. pdf.

[27] G. Wade, *Signal Coding and Processing*. Cambridge, U.K.: Cambridge Univ. Press, Sep. 1994.

[28] N. Mandagere, P. Zhou, M. A. Smith, J. Dong, and S. Uttamchandani, "Demystifying data deduplication," in *Proc. ACM/IFIP/USENIX Middleware Conf. Companion*, Dec. 2008, pp. 12–17.

[29] E. Keogh, S. Lonardi, and C. A. Ratanamahatana, "Towards parameter-free data mining," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2004, pp. 206–215.

[30] R. Cilibrasi and P. M. B. Vitanyi, "Clustering by compression," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1523–1545, Apr. 2005.

[31] A. N. Kolmogorov, "On tables of random numbers," *Theoretical Comput. Sci.*, vol. 207, no. 2, pp. 387–395, Nov. 1998.

[32] A. Macedonas, D. Besiris, G. Economou, and S. Fotopoulos, "Dictionary based color image retrieval," *J. Visual Commun. Image Representation*, vol. 19, no. 7, pp. 464–470, Oct. 2008.

[33] D. Cerra and M. Datcu, "A fast compression-based similarity measure with applications to content-based image retrieval," *J. Visual Commun. Image Representation*, vol. 207, no. 2, pp. 293–302, Feb. 2012.

[34] J. M. Lillo-Castellano, I. Mora-Jimenez, R. Santiago-Mozos, F. Chavarria-Asso, A. Cano-Gonzalez, and J. L. Rojo-Alvarez, "Symmetrical compression distance for arrhythmia discrimination in cloud-based big-data services," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 4, pp. 1253–1263, Jul. 2015.

[35] W. Xia, D. Feng, and L. Tian, "Combining deduplication and delta compression to achieve low-overhead data reduction on backup datasets," in *Proc. Data Compression Conf.*, Mar. 2014, pp. 203–212.

[36] M. Chen, S. Mao, Y. Zhang, and V. C. Leung, *Big Data: Related Technologies, Challenges and Future Prospects*. New York, NY, USA: Springer, 2014.

[37] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, "Analytics: The real-world use of big data," IBM Global Business Services, Somers, NY, USA, 2012.

[38] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (IoT): A vision, architectural elements, and future directions," *Elsevier Future Gener. Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660, 2013.

[39] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: Current state and future challenges," *IEEE Commun. Mag.*, vol. 49, no. 11, pp. 32–39, Nov. 2011.

[40] D. De Donno, L. Catarinucci, and L. Tarricone, "RAMSES: RFID augmented module for smart environmental sensing," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 7, pp. 1701–1708, Jul. 2014.

[41] H. Zou, Y. Yu, W. Tang, and H. M. Chen, "Improving I/O performance with adaptive data compression for big data applications," in *Proc. IEEE Int. Parallel Distrib. Process. Symp. Workshops*, 2014, pp. 1228–1237.

[42] C. Liu, K. Wu, and J. Pei, "An energy-efficient data collection framework for wireless sensor networks by exploiting spatiotemporal correlation," *IEEE Trans. Parallel Distrib. Syst.*, vol. 18, no. 7, pp. 1010–1023, Jul. 2007.

[43] C. Authors, "Big data spectrum," Infosys Limited, Bangalore, India, 2012.

[44] F. Marcelloni and M. Vecchio, "A simple algorithm for data compression in wireless sensor networks," *IEEE Commun. Lett.*, vol. 12, no. 6, pp. 411–413, Jun. 2008.

[45] G. Campobello, O. Giordano, A. Segreto, and S. Serrano, "Comparison of local lossless compression algorithms for wireless sensor networks," *J. Netw. Comput. Appl.*, vol. 47, pp. 23–31, 2015.

[46] L. Xiang, J. Luo, and C. Rosenberg, "Compressed data aggregation: Energy-efficient and high-fidelity data collection," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1722–1735, Dec. 2013.

[47] R. e ergelt, M. Vodel, and W. Hardt, "Energy efficient handling of big data in embedded,wireless sensor networks," in *Proc. IEEE Sensors Appl. Symp.*, Feb. 2014, pp. 53–58.

[48] T. Srisooksai, K. Keamarungsi, P. Lamsrichan, and K. Araki, "Practical data compression in wireless sensor networks: A survey," *J. Netw. Comput. Appl.*, vol. 35, no. 1, pp. 37–59, 2012.

[49] P. Jesus, C. Baquero, and P. S. Almeida, "A survey of distributed data aggregation algorithms," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 381–404, Mar. 2015.

[50] D. Takaishi, H. Nishiyama, N. Kato, and R. Miura, "Toward energy efficient big data gathering in densely distributed sensor networks," *IEEE Trans. Emerging Topics Comput.*, vol. 2, no. 3, pp. 388–397, Sep. 2014.

[51] J. Chen, W. Xu, S. He, Y. Sun, P. Thulasiraman, and X. S. Shen, "Utility-based asynchronous flow control algorithm for wireless sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 7, pp. 1116–1126, Sep. 2010.

[52] E. Zahavi, I. Keslassy, and A. Kolodny, "Distributed adaptive routing for big-data applications running on data center networks," in *Proc. 8th ACM/IEEE Symp. Arch. Netw. Commun. Syst.*, Oct. 2012, pp. 99–110.

[53] N. Pantazis, S. A. Nikolidakis, and D. D. Vergados, "Energy-efficient routing protocols in wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 551–591, May 2013.

[54] B. Bergler, C. Preschern, A. Reiter, and S. Kraxberger, "Cost-effective routing for a greener internet," *Proc. IEEE/ACM Int. Conf. Green Comput. Commun./Int. Conf. Cyber, Phys. Soc. Comput.*, Dec. 2010, pp. 276–283.

[55] A. Soran, F. M. Akdemir, and M. Yuksel, "Parallel routing on multi-core routers for big data transfers," in *Proc. CoNEXT Student Workhop*, Dec. 2013, pp. 35–38.

[56] F. A. Moghaddam, P. Lago, and P. Grosso, "Energy-efficient networking solutions in cloud-based environments: A systematic literature review," *ACM Comput. Surveys*, vol. 47, no. 4, pp. 64-1–64-32, May 2015.

[57] S. Ricciardi, D. Careglio, G. Santos-Boada, J. Sole-Pareta, U. Fiore, and F. Palmieri, "Towards an energy-aware internet: Modeling a cross-layer optimization approach," *Telecommun. Syst.*, vol. 52, no. 2, pp. 1247–1268, Feb. 2013.

[58] G. Wang, T. S. E. Ng, and A. Shaikh, "Programming your network at run-time for big data applications," in *Proc. ACM SIGCOMM 1st Workshop Hot Topics Softw. Defined Netw.*, Aug. 2012, pp. 103–108.

[59] Y. Wu, Z. Zhang, Y. Wu, Z. Zhang, C. Wu, C. Guo, Z. Li, and F. Lau, "Orchestrating bulk data transfers across geo-distributed datacenters," *IEEE Trans. Cloud Comput.*, accepted, doi: 10.1109/TCC.2015.2389842.

[60] R. Li, H. Harai, and H. Asaeda, "An aggregatable name-based routing for energy-efficient data sharing in big data era," *IEEE Access*, vol. 3, pp. 955–966, Jun. 2015, doi: 10.1109/ACCESS.2015.2448736.

[61] J. Ren, H. Wang, and L. Gao, "Research on low energy consumption of mass data download in smartphone," in *Proc. Int. Conf. Adv. Cloud Big Data*, Dec. 2013, pp. 20–26.

[62] Y. Wang, S. Su, A. Liu, and Z. Zhang, "Multiple bulk data transfers scheduling among datacenters," *Comput. Netw.*, vol. 68, pp. 123–137, Aug. 2014.

[63] R. T. Kaushik and M. Bhandarkar, "GreenHDFS: Towards an energy-conserving, storage-efficient, hybrid hadoop compute cluster," in *Proc. Int. Conf. Power Aware Comput. Syst.*, 2010, pp. 1–9.

[64] R. T. Kaushik, L. Cherkasova, R. Campbell, and K. Nahrstedt, "Lightning: Self-adaptive, energy-conserving, multi-zoned, commodity green cloud storage system," in *Proc. 19th ACM Int. Symp. High Performance Distrib. Comput.*, 2010, pp. 332–335.

[65] T. Bostoen, S. Mullender, and Y. Berbers, "Power-reduction techniques for data-center storage systems," *ACM Comput. Surveys*, vol. 45, no. 3, pp. 33-1–33-38, Jun. 2013.

[66] C. Negru, F. Pop, V. Cristea, N. Bessisy, and J. Li, "Energy efficient cloud storage service: Key issues and challenges," in *Proc. 4th Int. Conf. Emerging Intell. Data Web Technol.*, Sep. 2013, pp. 763–766.

[67] K. Mardamutu, R. A. L. A. Joon, J. Jegatheesan, V. Ponnusamy, and Y. Mei, "A critical analysis of green database system for energy efficiency and green computing," in *Proc. Interna. Sympo. Math. Sci. Comput. Research*, vol. 39, no. 1, Dec. 2013, pp. 11–17.

[68] S. Chaudhary, D. P. Murala, and V. Shrivastava, "Green database," *Global J. Bus. Manage. Inf. Technol.*, vol. 1, no. 2, pp. 105–111, 2011.

[69] L. Wu, R. J. Barker, M. A. Kim, and K. A. Ross, "Navigating big data with high-throughput, energy-efficient data partitioning," *ACM SIGARCH Comput. Arch. News*, vol. 41, no. 3, pp. 249–260, 2013.

[70] Q. Zhu, F. M. David, C. F. Devaraj, Z. Li, Y. Zhou, and P. Cao, "Reducing energy consumption of disk storage using power-aware cache management," in *Proc. IEE Softw.*, 2004, pp. 118–118.

[71] E. Pinheiro, R. Bianchini, and C. Dubnicki, "Exploiting redundancy to conserve energy in storage systems," in *Proc. ACM SIGMETRICS Perform. Eval. Rev.*, vol. 34, no. 1, 2006, pp. 15–26.

[72] L. Ganesh, H. Weatherspoon, M. Balakrishnan, and K. Birman, "Optimizing power consumption in large scale storage systems," presented at the 11th USENIX Workshop Hot Topics Operating Systems, San Diego, CA, USA, 2007.

[73] C. Liu, X. Qin, S. Kulkarni, C. Wang, S. Li, A. Manzanares, and S. Baskiyar, "Distributed energy-efficient scheduling for data-intensive applications with deadline constraints on data grids," in *Proc. IEEE Int. Performance, Comput. Commun. Conf.*, Dec. 2008, pp. 26–33.

[74] R. L. Villars, C. W. Olofson, and M. Eastwood, "Big data: What it is and why you should care," White Paper, International Data Corporation, Framingham, MA, USA, 2011.

[75] X. Gu, R. Hou, K. Zhang, L. Zhang, and W. Wang, "Application-driven energy-efficient architecture explorations for big data," in *Proc. 1st Workshop Arch. Syst. Big Data*, 2011, pp. 34–40.

[76] L. Liu, H. Wang, X. Liu, X. Jin, W. B. He, Q. B. Wang, and Y. Chen, "Greencloud: A new architecture for green data center," in *Proc. 6th Int. Conf. Ind. Session Autonomic Comput. Commun. Ind. Session*, 2009, pp. 29–38.

[77] Convey Computer Corporation. (2012). A big data computer architecture: The convey MX series. [Online]. Available: http://www.conveycomputer.com/files/6713/5266/3042/CONV-12-043.1MXdatac hureWeb.pdf.

[78] J. Leidel, J. Bolding, and G. Rogers, "Toward a scalable heterogeneous runtime system for the convey MX architecture," in *Proc. IEEE 27th Int. Parallel Distrib. Process. Symp. Workshops PhD Forum*. May 2013, pp. 1597–1606.

[79] High throughput computing data center architecture. [Online]. Available: http://www.huawei.com

[80] P. Dben, J. Schlachter, S. Y. Parishkrati, J. Augustine, C. Enz, K. Palem, and T. N. Palmer, "Opportunities for energy efficient computing: A study of inexact general purpose processors for high-performance and big-data applications," in *Proc. Design, Autom. Test Eur. Conf. Exhib.*, Mar. 2015, pp. 764–769.

[81] K. Kanoun, M. Ruggiero, D. Atienza, and M. van der Schaar, "Low power and scalable many-core architecture for big-data stream computing," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI*, Jul. 2014, pp. 468–473.

[82] A. Kejariwal, "Big data challenges: A program optimization perspective," in *Proc. Int. Conf. Cloud Green Comput.*, Nov. 2012, pp. 702–707.

[83] Y. Wang, B. Li, R. Luo, Y. Chen, N. Xu, and H. Yang, "Energy efficient neural networks for big data analytics," in *Proc. Design, Autom. Test Eur. Conf. Exhib.*, Mar. 2014, pp. 1–4.

[84] E. Park, J. Ahn, S. Hong, S. Yoo, and S. Lee, "Memory fast-forward: A low cost special function unit to enhance energy efficiency in GPU for big data processing," in *Proc. Design, Autom. Test Eur. Conf. Exhib.*, Mar. 2015, pp. 1341–1346.

[85] D. Gmach, Y. Chen, A. Shah, and J. Rolia, "Profiling sustainability of data centers," in *Proc. IEEE Int. Symp. Sustainable Syst. Technol.*, May 2010, pp. 1–6.

[86] G. B. Barone, D. Bottalico, V. Boccia, and L. Carracciuolo, "ECCO: An integrated solution for environment compatible computing systems," in *Proc. Int. Conf. Intell. Netw. Collaborative Syst.*, Sep., 2014, pp. 545–550.

[87] K. Schroder and W. Nebel, "Behavioral model for cloud aware load and power management," *Proc. Int. Workshop Hot Topics Cloud Serv.*, Apr. 2013, pp. 19–26.

[88] T. Mastelic, A. Oleksiak, H. Claussen, I. Brandic, J.-M. Pierson, and A. V. Vasilakos, "Cloud computing: Survey on energy efficiency," *Comput. Surveys*, vol. 47, no. 2, Jan. 2015.

[89] M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," *IEEE/ACM Trans. Netw.*, vol. 21, no. 5, pp. 1378–1391, Oct. 2013.

[90] Z. Wei and D. Q. Ren, "Review of energy aware big data computing measurements, benchmark methods and performance analysis," in *Proc. 23rd Int. Conf. Comput. Commun. Netw.*, Aug. 2014, pp. 1–4.

[91] Z. Wei and D. Q. Ren, "Review of energy aware big data computing measurements, benchmark methods and performance analysis," in *Proc. 23th Int. Conf. Comput. Commun. Netw.*, Aug. 2014, pp. 1–4.

[92] P. Brenner, R. Jansen, D. Go, and D. Thain, "Environmentally opportunistic computing: Transforming the data center for economic and environmental sustainability," in *Proc. Int. Green Comput. Conf.*, Aug. 2010, pp. 383–388.

[93] S. Ren, "Optimizing water efficiency in distributed data centers," in *Proc. IEEE 3rd Int. Conf. Cloud Green Comput.*, Sep. 2013, pp. 68–75.

[94] J. L. Berral, I. Goiri, R. Nou, F. Julia, J. Guitart, R. Gavalda, and J. Torres, "Towards energy-aware scheduling in data centers using machine learning," in *Proc. 1st Int. Conf. Energy-Efficient Comput. Netw.*, Apr. 2010, pp. 215–224.

[95] J.-C. Huet and I. E. Abbassi, "Green cloud computing modelling methodology," *Proc. IEEE/ACM 6th Int. Conf. Utility Cloud Comput.*, pp. 339–344, Dec. 2013.

[96] S. Pelley, D. Meisner, T. F. Wenisch, and J. W. VanGilder, "Understanding and abstracting total data center power," in *Proc. Workshop Energy Efficient Design*, 2009, pp. 1–6.

[97] G. Callou, P. Maciel, F. Magnani, J. Figueiredo *et al.*, "Estimating sustainability impact, total cost of ownership and dependability metrics on data center infrastructures," in *Proc. IEEE Int. Symp. Sustainable Syst. Technol.*, May 2011, pp. 1–6, May 2011.

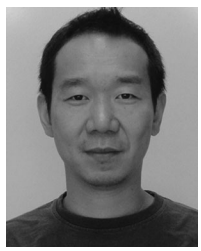
[98] B. Addis, D. Ardagna, A. Capone, and G. Carello, "Energy-aware joint management of networks and cloud infrastructures," *Comput. Netw., Int. J. Comput. Telecommun. Netw.*, vol. 70, pp. 75–95, Sep. 2014.

[99] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser, "Renewable and cooling aware workload management for sustainable data centers," in *Proc. 12th ACM SIGMET-*

RICS/PERFORMANCE Joint Int. Conf. Meas. Model. Comput. Syst., Jun. 2012, pp. 175–186.
[100] C. Li, Y. Hu, R. Zhou, M. Liu, L. Liu, J. Yuan, and T. Li, "Enabling datacenter servers to scale out economically and sustainably," *Proc. IEEE/ACM 46th Annu. Int. Symp. Microarchit.*, Dec. 2013, pp. 322–333.
[101] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.
[102] E. Dahlman, S. Parkvall, and J. Skold, *Hadoop: The Definitive Guide*, 4th ed. O'Reilly Media, Sebastopol, California, U.S., Apr. 2015.
[103] Q. Anderson, *Storm Real-Time Processing Cookbook Paperback*. Packt Publishing, Birmingham, U.K, Aug. 2013.
[104] P. T. Goetz and B. O'Neill, *Storm Blueprints: Patterns for Distributed Real-time Computation*. Packt Publishing, Birmingham, U.K., Apr. 2014.
[105] M. Hausenblas and J. Nadeau, "Apache drill: Interactive ad-hoc analysis at scale," *Big Data*, vol. 1, no. 2, pp. 100–104, 2013.
[106] D. Cavdar, L. Y. Chen, and F. Alagoz, "Green MapReduce for heterogeneous data centers," in *Proc. IEEE Global Commun. Conf.*, 2014, pp. 1120–1126.
[107] Í. Goiri, K. Le, T. D. Nguyen, J. Guitart, J. Torres, and R. Bianchini, "Greenhadoop: Leveraging green energy in data-processing frameworks," in *Proc. 7th ACM Eur. Conf. Comput. Syst.*, 2012, pp. 57–70.
[108] G. Cavallaro, M. Riedel, J. Benediktsson, M. Goetz, T. Runarsson, K. Jonasson, and T. Lippert, "Smart data analytics methods for remote sensing applications," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2014, pp. 1405–1408.
[109] L. Mashayekhy, M. M. Nejad, D. Grosu, Q. Zhang, and W. Shi, "Energy-aware scheduling of MapReduce jobs," in *Proc. IEEE Int. Congr. Big Data*, Jul. 2014, pp. 32–39.
[110] R. T. Kaushik and K. Nahrstedt, "T^*: A data-centric cooling energy costs reduction approach for big data analytics cloud," in *Proc. Int. Conf. High Perform. Comput., Netw. Storage Anal.*, Nov. 2012, pp. 1–11.
[111] S. Roy and S. Gupta, "The green cloud effective framework: An environment friendly approach reducing CO2 level," in *Proc. 1st Int. Conf. Non Conventional Energy*, Jan. 2014, pp. 233–236.
[112] L. D. Lathauwera and J. Vandewalleb, "Dimensionality reduction in higher-order signal processing and rank-$(r1, r2, , rn)$ reduction in multilinear algebra," *Linear Algebra Appl.*, vol. 391, pp. 31–55, Nov. 2004.
[113] H. Wang and N. Ahuja, "Compact representation of multidimensional data using tensor rank-one decomposition," in *Proc. 17th Int. Conf. Pattern Recog.*, vol. 1, Aug. 2004, pp. 44–47.
[114] H. Wang and N. Ahuja, "A tensor approximation approach to dimensionality reduction," *Int. J. Comput. Vis.*, vol. 76, no. 3, pp. 217–229, Mar. 2008.
[115] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos, "Tag recommendations based on tensor dimensionality reduction," in *Proc. ACM Conf. Recommender Syst.*, Oct. 2008, pp. 203–212.
[116] E. Acar, R. J. Harrison, F. Olken, O. Alter, M. Helal, L. Omberg, B. Bader, A. Kennedy, H. Park, Z. Bai, D. Kim, R. Plemmons, G. Beylkin, T. Kolda, S. Ragnarsson, L. DeLathauwer, J. Langou, S. P. Ponnapalli, I. Dhillon, L.-H. Lim, J. R. Ramanujam, C. Ding, M. Mahoney, J. Raynolds, L. Elden, C. Martin, P. Regalia, P. Drineas, M. Mohlenkamp, P. S. Sadayappan, C. Faloutsos, J. Morton, B. Savas, S. Friedland, L. Mullin, and C. V. Loan. (2009, May). Future directions in tensor-based computation and modeling. presented at the NSF Workshop Future Directions Tensor-Based Computation Modeling. [Online]. Available: http://www.cs. cornell. edu/cv/tenwork/finalreport. pdf.
[117] H. Lua, K. N. Plataniotisb, and A. N. Venetsanopoulosb, "A survey of multilinear subspace learning for tensor data," *Pattern Recog.*, vol. 44, no. 7, pp. 1540–1551, Jul. 2011.
[118] S. W. Jun, K. E. Fleming, M. Adlery, and J. Emery, "ZIP-IO: Architecture for application-specific compression of big data," in *Proc. Int. Conf. Field-Programmable Technol.*, Dec. 2012, pp. 1–11.
[119] C. Bi, K. Ono, K.-L. Ma, H. Wu, and T. Imamura, "Proper orthogonal decomposition based parallel compression for visualizing big data on the k computer," in *Proc. IEEE Symp. Large-Scale Data Anal. Visual.*, Oct. 2013, pp. 121–122.
[120] L. Wang, H. Zhong, H. Zhong, R. Ranjan, A. Zomaya, and P. Liu, "Estimating the statistical characteristics of remote sensing big data in the wavelet transform domain," *IEEE Trans. Emerging Topics Comput.*, vol. 2, no. 3, pp. 324–337, Sep. 2014.
[121] F. Marcelloni and M. Vecchio, "An efficient lossless compression algorithm for tiny nodes of monitoring wireless sensor networks," *Comput. J.*, vol. 52, no. 8, pp. 969–987, 2009.
[122] M. G. Patterson, "What is energy efficiency? Concepts, indicators, and methodological issues," *Energy Policy*, vol. 24, no. 5, pp. 377–390, May 1996.
[123] "Environmental engineering (EE) energy efficiency of wireless access network equipment," ETSI TS 102 706 V1.1.1, Aug. 2009.
[124] F. Richter, A. Fehske, and G. Fettweis, "Energy efficiency aspects of base station deployment strategies for cellular networks," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2009, pp. 1–5.
[125] D. Tsirogiannis, S. Harizopoulos, and M. A. Shah, "Analyzing the energy efficiency of a database server," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, Jun. 2010, pp. 231–242.
[126] T. Q. S. Quek, W. C. Cheung, and M. Kountouris, "Energy efficiency analysis of two-tier heterogeneous networks," in *Proc. Eur. Wireless Conf.*, Apr. 2011, pp. 1–5.
[127] H. T. Odum, "Self-organization, transformity, and information," *Science*, vol. 242, no. 4, pp. 1132–1139, 1988.
[128] H. T. Odum, *Environment Accounting: Emergy and Environment Decision Making*. New York, NY, USA: Wiley, 1996.
[129] W. E. Rees, "Ecological footprints and appropriated carrying capacity: What urban economics leaves out," *Environ. Urbanization*, vol. 4, no. 2, pp. 121–130, Oct. 1992.
[130] N. Fiala, "Measuring sustainability: Why the ecological footprint is bad economics and bad environmental science," *Ecological Econ.*, vol. 67, no. 4, pp. 519–525, Nov. 2008.
[131] L. Gu, D. Zeng, P. Li, and S. Guo, "Cost minimization for big data processing in Geo-Distributed data centers," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 314–323, Sep. 2014.
[132] J. Cao and M. Yang, "Energy Internet—Towards smart grid 2.0," in *Proc. 4th Int. Conf. Netw. Distrib. Comput.*, 2013, pp. 105–110.

Jinsong Wu (SM'11) received his Ph.D. in Department of Electrical and Computer Engineering, Queen's University at Kingston, Canada in 2006, and is currently an Associate Professor in the Department of Electrical Engineering, Universidad de Chile, Santiago, Chile. He is the Founder and Founding Chair of the Technical Committee on Green Communications and Computing, IEEE Communications Society. He is also the Cofounder and Founding Vice Chair of the Technical SubCommittee on Big Data, IEEE Communications Society. He is the Founder and Editor of Series on Green Communication and Computing Networks, *IEEE Communications Magazine*. He is an Editor of theIEEE J. SELECT. AREAS ON COMMUNICATIONS Series on Green Communications and Networking. He has served as the leading Guest Editor of Special Issue on Green Communications, Computing, and Systems inIEEE SYSTEMS JOURNAL, Associate Editor of Special Section on Big Data for Green Communications and Computing in IEEE ACCESS. He was the leading Editor and a coauthor of the comprehensive book, entitled "Green Communications: Theoretical Fundamentals, Algorithms, and Applications," (Boca Raton, FL, USA: CRC Press, Sep. 2012).

Prof. Wu was the leading General Chair in the 2013 IEEE International Conference on Green Computing and Communications.



Song Guo (SM'11) received the Ph.D. degree in computer science from the University of Ottawa, Canada.

He is currently a full Professor at the University of Aizu, Aizuwakamatsu, Japan. His research interests include areas of wireless network, cloud computing, big data, and cyber-physical system. He has authored/edited 7 books and more than 300 papers in refereed journals and conferences in these areas. He serves/served in editorial boards of IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING, *IEEE Communications Magazine*, *Wireless Networks*, *Wireless Communications and Mobile Computing*, and many other major journals.

Dr. Guo has been the General/Program Chair or in organizing committees of numerous international conferences. He is a senior member of the ACM and an IEEE Communications Society Distinguished Lecturer.

Jie Li (M'94–SM'04) received the B.E. degree in computer science from Zhejiang University, Hangzhou, China, the M.E. degree in electronic engineering and communication systems from the China Academy of Posts and Telecommunications, Beijing, China, and the Dr. Eng. degree from the University of Electro-Communications, Tokyo, Japan.

He is with the Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, Japan, where he is a Professor. He has been a Visiting Professor in Yale University, New Haven, CT, USA, Inria, France. His current research interests include mobile distributed computing and networking, big data and cloud computing, IoT, OS, and modeling and performance evaluation of information systems.

Dr. Li is a senior member of the ACM and a member of the Information Processing Society of Japan (IPSJ). He is the Chair of Technical Subcommittee on Big Data, IEEE Communications Society. He has served as a Secretary for Study Group on System Evaluation of IPSJ and on several editorial boards for the international Journals, and on Steering Committees of the SIG of System Evaluation of the IPSJ, the SIG of DataBase System of the IPSJ, and the SIG of MoBiLe computing and ubiquitous communications of the IPSJ. He has also served on the program committees for several international conferences such as the IEEE INFOCOM, IEEE GLOBECOM, and IEEE MASS.

Deze Zeng (M'14) received the B.S. degree from the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China in 2007, and the M.S. and Ph.D. degrees in computer science from the University of Aizu, Aizu-Wakamatsu, Japan, in 2009 and 2013, respectively.

He is currently an Associate Professor in the School of Computer Science, China University of Geosciences, China. His current research interests include cloud computing, software-defined sensor networks, data center networking, and networking protocol design and analysis.