This is a postprint version of the following published document:

Zhihan Lv; Houbing Song; Pablo Basanta-Val; Anthony Steed; Minho Jo. *Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics*. IEEE Transactions on Industrial Informatics (2017), 13(4), 1891 – 1899.

Doi: http://dx.doi/10.1109/TII.2017.2650204

© 2017 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics

Zhihan Lv, Houbing Song, *Senior Member, IEEE*, Pablo Basanta-Val, Anthony Steed,
and Minho Jo, *Senior Member, IEEE*

The term *big data* occurs more frequently now than ever before. A large number of fields and subjects, ranging from everyday life to traditional research fields (i.e., geography and transportation, biology and chemistry, medicine and rehabilitation), involve big data problems. The popularizing of various types of network has diversified types, issues, and solutions for big data more than ever be-fore. In this paper, we review recent research in data types, storage models, privacy, data security, analysis methods, and applications related to network big data. Finally, we summarize the challenges and development of big data to predict current and future trends.

I. Introduction

WE ARE living in an era of big data—an age characterized by fast collection of ubiquitous information. Big data incorporates endless amount of information [56]. In many industries, it is growing, providing a means to improve and streamline business. Many fields and sectors, ranging from economic and business activities to public administration, from national security to scientific research in many areas, are involved in big data problems [59]. Big data has changed the world in terms of predicting customer behavior. The birth of big data cannot avoid mentioning another current popular term—*social networks*—and the relation between the two is obvious, yet complicated.

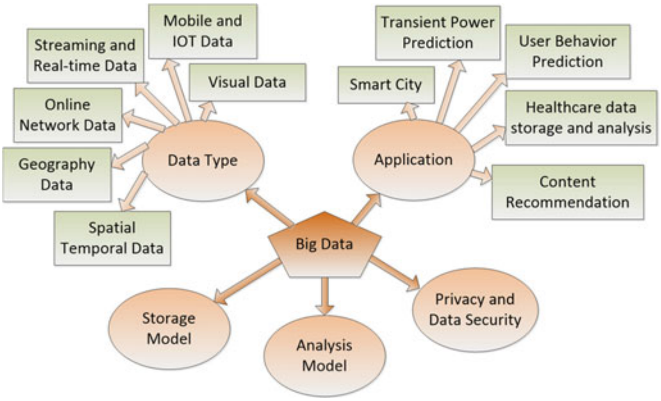

Fig. 1. Classification of big data literature.

Big data and social networks are interdependent, because most of today's data are generated from social networking sites, but big data is not always useful. The actual challenge of big data is not in collecting it, but in managing it as well as making sense of it [63]. When we work on big data, it is crucial to determine whether the benefits outweigh the costs of storage and maintenance. Several tools are being designed to better understand the role of huge amount of data in improving business. Researchers and practitioners are trying to look into the future of big data to extract more benefits

Big data is used in several research areas related to healthcare, location-based services, satellite information usage, online advertising, and retail marketing. In the coming years, the Internet of Things (IoT) will increase the amount of data in the world, and an exponential rise in big data will be seen [1]. There have been a few review papers on big data in general [2], [18] [26], [56], [59], [63], [66], [67] and in diverse specialized fields e.g., biology [61], healthcare [62], geography [65], and the Internet [60]. Unlike previous reviews in the literature, this paper looks at recent advances in big data from a different view and with different classifications i.e., data types, storage models, analysis models, privacy, data security, and applications, as shown in Fig. 1.

The rest of the paper is organized as follows. Section II presents data types. Section III gives the storage model. Section IV discusses the privacy and data security. Sections V and VI present various big data analysis methods and applications, respectively. Section VII concludes this paper with future research topics.

II. Data Types

The era of big data has produced a variety of datasets from different sources in different domains. These datasets consist of multiple modalities, each of which has a different representation, distribution, scale, and density. How to unlock the power of knowledge from disparate datasets is of paramount importance in big data research, essentially distinguishing big data from traditional data mining tasks [45].

A. Online Network Data

One of the main focuses of network big data is online social network (OSN) data, such as Facebook [58] and Second Life [68]. This focus has expanded with developments in the data analysis.

Many studies have been performed with OSNs using knowledge about representative characteristics at the macrolevel, for instance, small world features. However, factors for the features of potential microprocesses are not well represented in these studies. The simplification of math expo is in general viewed as a model by scholars. One study adheres to an additional strategy that results in the microscopic process of knowledge acting in accordance with real online actions [2]. Knowledge in this respect can be applied to carry out the choice of easily handled mathematical models in the generation and transformation of online networks.

Another study proposes a method for judging the intensity and category of social relations by relying on their spatial–temporal interrelations [5]. The result suggests that the method discussed in the paper is capable of successfully recognizing different social relations between different substances. With the nonstop development of social networks, more and more sociologists have become involved in the study of big data.

B. Mobile and IoT Data

Another trend in network big data is the analysis of mobile and IoT data.

With the development of 5G technology, converged mobile networks have resulted in significant improvements in machine-to-machine communications performance. Integrated mobile webs share unlicensed spectrum bands in cellulite networks, such as long-term evolution-advanced, by using cognitive radio technology. This network generates large volumes of data, compared to former mobile networks [3].

In addition to the increased volume of mobile data, the IoT also generates large amount of data in this new context. Despite this large volume of data, the sensing elements of wireless body area networks (WBANs) to a certain degree restrict power use [25], [54]. The majority of researchers attach importance to energy efficiency in media access control (MAC) agreements in lengthening the lifetime of the sensors. One study addresses the recognition of classification of power consumption attacks in MAC agreements in WBANs. It describes the straightforward operation of the attacks, resulting in power consumption in a variety of MAC agreements [4]. This work is a good reference for research on the power efficiency of MAC agreements in WBANs of the future.

Understanding the connection and interaction of mobile OSN data has been continuously broadened, although network big data in the IoT is a relatively new field The data structure of the analysis is more likely to be not only structured query language (NoSQL), which is adopted by many IoT systems. Some studies design the expected functions of a big loader and a convenient loading NoSQL system. The system allows the standard conceptual program to be loaded and lets the standard sources from which the data are supposed to be collected meet its requirements; finally, this study provides feasible strategies for the choice of the NoSQL system where the conceptual program can be arranged well [6].

C. Geography Data

OSN data will soon include geographic data along with OSN interaction, for example, geo-tag real-time geographic data [65]. Location-based data will soon expand beyond terrain.

One study addresses the gauntlets of major forms of technology for three-dimensional (3-D) interaction and volume-rendering technology on the basis of graphics processing unit (GPU) technology. This work explores visual software for the hydrological environment based on data orientation. In addition, it produces ocean plans, contour mapping of surfaces, element field mapping, and dynamic simulation of the existing field [7]. To better present features in space and achieve real-time upgrading of a large amount of hydrological environment data, the study constructs nodes on the spot for the control of geometry to achieve dynamic mapping of high properties.

With full-speed development of the 3-D digital city, research has shifted from model establishment of the 3-D city and the setting up of geo-databases to 3-D geo-database services and maintenance. There is also a paper putting forward an event-driven spatiotemporal database model in which 3-D city models from the past and present are connected to sentence meaning and representation of the state, touched off by changing events defined in advance [8]. In addition, the dissertation also presents a homologous dynamic renewal method taking an adaptive matching algorithm as the premise. The objective, depending on the compound matching of sentence-meaning, properties, and positions in space, is to carry out dynamic renewal for complex 3-D city models without being controlled by others.

D. Spatial Temporal Data

Accompanied by online streaming services, network big data changes from simple OSN data to spatial–temporal data. In recent years, the volume of available data from space has increased substantially. For example, in November 2013, NASA announced the release of hundreds of terabytes of Earth remote-sensing datasets.

Data are classified in many categories on the basis of features and differences. Since the differences in data determine the success of the analysis, they play an essential role. Different features are also applied to search for the same features. Some studies attach importance to time-changing data and data with time sequences. With respect to network big data, some same-feature searching methods for time-changing data were discussed and explored [12].

Data in large databases can be retrieved by data mining. In the case of time-changing data, when time becomes connected, the data are mined in terms of both time and space. The exploration of data mining in terms of both time and space has had a great influence on the study of data derived from mobile devices [9]. Reality mining is the exploration of social behavior according to data retrieved from mobile phones. That is to say, it depends on the data collected by sensors in mobile phones, security cameras, radio frequency identification (RFID) readers, and so on. NetViz, Gephi, and Weka have been applied in the conversion and analysis of Facebook data.

Spatial big data, in general, is involved in vector data, raster data, and network data. The difficulty with using databases from the perspective of space is that there are many obstacles at the collateral level. One study reviewed the data algorithms that are in general use, in particular for cases where the number of satellite images continuously increases [10]. It proposes a system under which Hadoop is employed to realize a MapReduce model, thus making it possible to enhance the category of large-scale remote sensing images and magnify the power of big data as applied to space. The research holds that a novel system architecture is capable of providing support for alternate data and data in space [11]. This is shown as follows:

1) the study represents an analysis framework for expanding data,
2) it illustrates novel structures and algorithms that leverage modern hardware, such as solid state drives, and
3) it extends database systems in space, thus lending support for an evaluation in space through three novel constituents.

E. Streaming and Real-Time Data

Accompanied by the rise in online streaming services, network big data has evolved from spatial–temporal data to real-time spatial–temporal data.

Network surveys in general require ongoing data analysis owing to constant renewal of reports and statistics over large-capacity data streams. In one study, researchers introduced DBStream, a system based on SQL that relies on surveys for the continuous data analysis [15]. They also discuss the respective properties of DBStream and the collateral data handling engine Spark. It is suggested that, on some occasions, a single DBStream node is capable of surpassing a set of ten Spark nodes due to the renewed network survey capacity.

The epoch of big data has begun, and much of the data are used to analyze the risks of a variety of industrial applications. There are technological trials in the collection of big data in a complex indoor industrial environment. Indoor wireless sensor network (WSN) technology is capable of overcoming such restrictions by gathering big data obtained from source nodes. The data are transferred to a data center, at present. In the study, representative housing, bureaus, and manufacturing environments were selected [13]. Through analysis of tested data, it is possible to obtain signal transferring features of an indoor WSN. On the basis of these features, a big data collection algorithm that relies on an indoor WSN was put forward for the analysis of industrial risk processing.

City traffic also changes in real time. Traffic data are regarded as worthwhile resources in networks of vehicle. Highlighting the significance of a survey of big data, an effective framework was put forward for current-time network data in vehicle networks [14]. The system, in fact, reflects the newest trends in big-data paradigms. The framework put forward is composed of concentrated data memory principles for a series of processes, and dispersed data memory principles for stream processing in real time.

The present big data streams from social networks and other associated sensor networks display the potential for relying on each other, thus enabling a special approach to the analysis of extended figures. Data from these figures are often gathered from data servers in various geographic locations, making it appropriate for dispersed handling in the cloud. While many measurements for large-scale immobile figure analysis have been brought forward, providing current-time analysis of the dynamics of social correlations requires novel methods that leverage increased stream handling and figure analysis in flexible cloud environments. Agnihotri and Sharma [16] put forward feasible measurement that depends on a stream handling engine referred to as Floe; on top of this framework, it is possible to implement current-time data handling and figure renewal to carry out analysis of figures with low delay in large-scale, constantly changing social networks. The scope contains multiple fields involving supervision, antiterrorist applications, and public health supervision.

Currently, space-borne sensors channel nearly constant streams of Earth-survey datasets. These tremendous multimodal streams increase at a rapid rate, presently reaching several petabytes of satellite files. An extended platform for both geography and space was devised, developed, and assessed for online and current-time gains of worthwhile content from big Earth-survey data [15]. The key features of the analysis platform are the Rasdaman Array Database Management System for big raster data memory, and the Open Geospatial Consortium Web Coverage Processing Service for data inquiry. The system was verified for self-acting handling of high resolving-power satellite data, as well as for major geographic and dimensional, environmental, agricultural, and water engineering use, e.g., precise agriculture, water quality control, and land-cover mapping.

Besides, there are some pieces of work that may inspire research in their related fields [69]–[74].

F. Visual Data

In the era of big data, ever increasing amount of image data has posed significant challenges to modern image analysis and retrieval. Wu *et al.* [47] proposed weakly semisupervised deep learning for the multilabel image annotation (WeSed) approach, which was inspired by recent advances in deep learning research. In WeSed, a novel weakly weighted pairwise ranking loss is effectively utilized to handle weakly labeled images, while a triplet similarity loss is employed to harness unlabeled images. WeSed enables users to train a deep convolutional neural network (CNN) with images from social networks, where images are either only weakly labeled, have several labels, or are unlabeled. An efficient algorithm was also designed to sample high-quality image triplets from large image datasets to fine-tune the CNN.

It is of great importance to efficiently and effectively index images with semantic keywords, particularly when confronted with the fast-evolving properties of the web. Yang *et al*. [48] proposed an unsupervised hashing approach, namely, robust discrete hashing (RDSH), to facilitate large-scale semantic indexing of image data. Specifically, RDSH simultaneously learns discrete binary codes as well as robust hash functions within a unified model. Extensive experiments were conducted on various real-world image datasets to show its superiority to state-of-the-art approaches in large-scale semantic indexing.

G. *Challenges in Data*

Each different data domain raises particular challenges that, properly addressed, may have an important impact on next-generation big data systems. In the first place, online network data are still waiting for better models, with increased support from sociologists. Mobile data and the IoT, which are generating large amount of data, would benefit from the adoption of a big data infrastructure able to store and process information in current IoT infrastructures. As for geography data, a major trend seems to be to offer efficient integration among geographic data with records from the OSN domain, demanding efficient infrastructures able to meet the speed requirements typical of these domains. One challenge imposed by spatial data is the definition of proper mining algorithms that can be applied to special data; those algorithms could benefit from more efficient time-changing data. Besides, streaming and real-time data challenge current infrastructures, transforming current offline applications into online ecosystems, thus requiring the development of new algorithms that take into account offline and online data. Finally, the ever-increasing amount of image data challenges current learning algorithms to extract information, and also demands new algorithms to semantically classify and index images.

In all these cases, all approaches would benefit from increased performance in big data infrastructures. By increasing efficiency in the different data domains, one can see how the amount of functionality improves; this is of special interest for the next generation of big data systems, which should integrate online and offline data efficiently

III. Storage Model

In the era of big data, the most difficult problems that remain to be solved are how to efficiently deal with large quantities and varieties of data. There are many analytical theories and models. In this section, recent discoveries in big data storage and analysis models are surveyed.

The acquisition of voluminous data depends on a variety of users and devices, as well as powerful data centers to store and process the data. For this reason, establishing an unimpeded network infrastructure is urgently needed; this infrastructure would make it possible to gather geologically distributed and rapidly generated data and send them to data centers for end users. In one study, participants witnessed the various challenges in establishing such a network framework [18]. This study presents the components of the network that must be established: networks relating to the original data, bridges to connect and transmit them to data centers, and intricate networks distributed within the data centers, as well as independent data centers.

Another study mainly addressed possible issues when using big data in specific locations, taking social network sites as one example [19]. The survey shows that users do not choose data arbitrarily when using this data network. On the contrary, most consciously try to find a particular site from data centers. This shows that one can identify the necessary data within a vast and complicated network.

Load counterpoising technologies, such as work usurping, play a major role in dispersed assignment-arranging systems; these systems possess various kinds of managers who determine the arrangement result [20] and facilitate work usurping by applying both contributed and borrowed alignments. Tasks are divided into several queues according to size and site. Skills are organized in MATRIX, a dispersed mission manager for mission computing. The researchers leveraged dispersed key-value memory to manage and measure the mission metadata, mission reliability, and data locality.

A structure that has a dispersed memory level local to the compute nodes was suggested [21]. This level is in charge of the majority of input/output (I/O) handlings, and economizes excessive amount of data movement between compute and memory origins. The study describes and implements a system antetype of this structure, which requires the FusionFS dispersed document system to sustain metadata-concentrated and write-concentrated operation, both of which play an essential role in I/O representation of scientific utilizations. FusionFS was developed and assessed based on 16K compute nodes of an IBM Blue Gene/P supercomputer; this suggests an order of scale-representation enhancement over other document systems, such as general parallel file system, parallel virtual file system, and Hadoop distributed file system (HDFS).

Another study shows the collection of dispersed key-value memory mechanisms in clouds and supercomputers [43]. Specifically, zero hop distributed hash table (ZHT), a zero-hop dispersed key-value memory system, is shown rearranging the requests of advanced computing systems. This system is meant to establish an alternative to other systems, such as collateral and dispersed document systems, dispersed working controlling systems, and collateral planning systems. The study also mentions systems that have incorporated ZHT in real applications, namely, FusionFS (a distributed file system), IStore (a storage system with erasure coding), MATRIX (distributed scheduling), Slurm++ (distributed high performance computing job launch), and Fabriq (distributed message queue management). Furthermore, it was also shown in another paper that certain superior computing systems are apt to adopt ZHT for their requests [22].

The absorbing ability with respect to data networks was discussed in other papers [23]. The research explored immediate aerial reception to effect immediate receiving re-establishment using two procedure line programs to initiate a multicast tree for a wireless multihop meshwork. Based on different means, the research aims at making every site possess a recognition capability, so all of them have the ability to absorb the smallest conveying power through perceiving, studying, behaving, and determining.

Peer-to-peer (P2P) overlay networks were also used in 3-D geographic big data searches [31]. Major achievements include drawing geographic and virtual space based on P2P coverage of the network space; and the spaces are classified by the quaternary tree approach. The geographic code is determined by hash value, being applied to index the user list, landform information, and the message of the model target. It is possible to devise and realize the sharing of data based on enhancement of the Kademlia meshwork pattern. In this pattern, an XOR algorithm is applied to calculate the range of cyberspace. The pattern, to a certain degree, enhances the hit rate of 3-D geographic data exploration under a P2P coverage meshwork.

As for storage models, the main challenge seems to be to efficiently deal with larger amount of data. In most cases, the lack of ultrascalable solutions is hindering the processing of other data sources, causing inefficiency. The challenge is to build a more scalable big data technology able to offer data gathering and distribution among nodes geographically dispersed across the world. Any improvement in this challenge is to going to have a direct impact on a variety of developed applications, which otherwise may suffer from technological constraints.

IV. Privacy and Data Security

Because of the scope of big data, safety and privacy protection is a crucial problem [53]. There may be risks of privacy violations at each step. There are many methods for privacy protection, e.g., encryption.

The popularity of big data depends on a complete understanding of the safety problems inherent within the system. Safety is a new concern, and this paper mainly introduces the concept of privacy using new problems, and focuses on efficiency and privacy protection. This study specializes in the structure of big data analytics, demonstrating the requirements for privacy protection; in addition, it explains the safety protection cosine similarity agreement in data mining and requirements.

In body area networks, wearable sensors collect data that are often sensitive and must be protected. Compared to former methods, this method provides more reliable and available privacy protection. The experiments prove that this method has sufficient privacy protection, even when hackers have adequate knowledge of the system.

System problems that are related to network intrusion forecasts are discussed. The paper focuses on problems dealing with big data categories, employing the techniques of geometric representation learning and modern networks. In particular, to overcome network traffic problems, this paper focuses on the problems associated with the technologies of supervised learning, representation learning, machine life-long learning, and big data (Hive and Could, etc.).

An International Data Corporation survey showed that, in 2011, 1.8 trillion gigabytes of data were created and copied, and that amount is duplicated every two years. In the coming decade, the total amount of data center-managed information will be 50 times larger; however, professional IT staff will only grow by 1.5 times. Conventional tools are not able to process and deal with the information contained in this amount of data, nor can they ensure security.

Because of the competitive market, customer management has become a means to achieving competitive advantage. The present churn prediction models do not work efficiently in a big data environment. Additionally, the deciders often face inaccurate management. In response to these challenges, a clustering algorithm referred to as semantic-driven was put forward. The experiment results showed that semantic-driven subtractive clustering method (SDSCM) possesses stronger semantic strength than the subtractive and fuzzy means. Thus, based on the Hadoop MapReduce framework, the SDSCM algorithm was realized.

Current state-of-the-art applications would benefit from definition of a clear hierarchy for data security and privacy, with common off-the-shelf policies designed for big data. Those policies should be able to efficiently deal with the definition of what is allowed and what is not, and should take into account diverse scenarios, ranging from body area networks (BANs) to large data servers that process larger amount of data. The accomplishment of this goal in an organized way will promote the adoption of high-quality big data systems and applications.

V. Analysis Methods

At present, the methods used for big data analysis are MapReduce-related. For data control in the past, instruments for analyzing data were insufficient in depot and exploring systems. The models used by big data researchers are usually inspired by mathematical ease of exposition [60]. By virtue of the essence of big data, it is memorized in a dispersed document system framework. Hadoop and HDFS by Apache are extensively applied in memorizing and controlling big data. The exploration of big data is fraught with obstacles, since it is related to large dispersed document systems that are supposedly featured by mistake endurance, agility, and the ability to be extended [65]. MapReduce is extensively applied for productivity exploration of big data [31]. Conventional database management system technologies including Joins and Searching and others, such as drawing exploration, are utilized in the division and integration of big data [27]. These technologies are applied in MapReduce.

Arora and Chana [28] suggest a variety of approaches to address the issues that stem from a MapReduce structure over a Hadoop distributed file system. MapReduce is the technology that utilizes document searching for drawing, classifying, shuffling and decreasing. Researchers have performed research on MapReduce technologies for big data applications based on HDFS [28]. One paper also studied MapReduce in a mobile environment [29].

Another study presents a security framework for large-space multimedia file [30]. A hybrid safety cloud memory structure is proposed that relies on the IoTs. It applied the idea of multimedia defense on the basis of role-visiting domination. In addition, the study also made use of a program that takes the integration of the multimedia data state and role-visiting domination as the premise. The IoTs is applied to assess whether circuits are associated and whether the equipment is running naturally to improve visiting efficiency, which ensures the safety of multimedia files

Many researchers have attempted to produce systematic ways of applying a wide spectrum of advanced machine learning (ML) programs to industrial-scale problems. Xing *et al*. proposed a general-purpose framework, Petuum, which systematically addresses data- and model-parallel challenges in large-scale ML by observing that many ML programs are fundamentally optimization-centric and admit error-tolerant, iterative-convergent algorithmic solutions [46]. This presents unique opportunities for an integrative system design, such as bounded-error network synchronization and dynamic scheduling based on an ML program structure.

Here, one of the cornerstones is the MapReduce model that offers support to most analyses. The MapReduce model was designed to process offline data in a batch-processing engine; however, next-generation analytics, which also have online requirements, should benefit from a lightweight version of MapReduce and the implantation of distributed stream processing engines.

VI. Applications

As data and interactions are generated in every form of human behavior, big data is used in almost all aspects of life. Big data increasingly benefit both research and industrial fields such as healthcare, financial services, and commercial recommendations. Big data is used primarily to predict certain messages, such as transient power [34] and stock prices [35].

A. Transient Power Prediction

The prediction of transient power is valid in both distributed and streaming data. ML was used in the study. In the classifier cultivation stage, researchers regard the tremendous amount of data from the past as a dispersed study target, and establish evaluation principles regularly. Zhiwei *et al*. [34] designed a naive Bayes-category approach based on MapReduce handling, creating a map-and-decrease procedures method for calculating the chance rate of being tested in advance and the chance rate for conditions in dispersed means.

B. User Behavior Prediction

Many of the network big data predictions are based on data from OSNs. Big data is used for predictions based on ranked data, such as elections, car performance, and other areas in business and politics. One study discussed modeling and analysis approaches to democracy, as well as various cases of big data from elections; scenarios in established democracies such as the United States and Canada, and new democracies such as Tunisia, were studied [36]. Another study gathered and explored user practices on Facebook. The model is capable of arranging entities with effectiveness and efficiency (for example, presidential candidates, specialized sport groups, and musical bands) according to their popularity [37].

C. Healthcare Data Storage and Analysis

Big data in health and biology to tackle the challenges in new models [61], [62] is becoming significant One study introduced two uses of mHealth, which gathers electronic medical records that are used for health services terminals [38]. One is a blended system that enhances the user experience in high-pressured oxygen halls using virtual reality (VR) glasses, which creates the feeling of being inside it. The other is a sound interaction game that is used by patients as a possible measurement for supplementary recovery tools. It is possible to analyze recordings of the sounds made by patients to assess long-term recovery results and further forecast the recovery process.

D. Content Recommendation

One study presents a movie recommendation system based on scores provided by users. In view of the movie evaluation system, the impacts of access control and multimedia security are analyzed, and a secure hybrid cloud storage architecture is presented [41]. Mobile-edge computing technology is used in the public cloud, which guarantees high-efficiency requirements for the transmission of multimedia content [55]. The processes of the system, including registration, user login, role assignment, data encryption, and data decryption, are also described.

Personalized travel sequence recommendation was proposed in another study [49], which uses travelogues, community-contributed photos, and heterogeneous metadata (e.g., tags, geo-location, and date taken) associated with the photos. This method is not only personalized to user travel interests but is also able to recommend a travel sequence rather than individual points of interest (POIs). To recommend personalized POI sequences, first well-known routes are ranked according to similarity between user packages and route packages. Then, top-ranked routes are further optimized by similar user travel records. Representative images with viewpoints and seasonal diversity of POIs are shown to offer a more comprehensive impression.

E. Smart City

A 3-D Shenzhen city web platform based on a network virtual reality geographic information system (GIS) was put forward [39]. A 3-D worldwide browser is applied to load different kinds of required data from a city, such as 3-D construction model data, inhabitants' messages, and traffic data from the past and present. These data are used to analyze and visualize city information on a 3-D platform. A large number of messages are capable of being visualized on this platform, and a navigational project, taking the GIS as the premise, makes it possible to obtain a variety of data sources that are securable.

The enhancive requirement for fluidity has resulted in great changes in fundamental facilities in transportation [42]. Possessing certain features, such as a large scale, diversified foreseeability, and timeliness, city traffic data represent the scope of big data [40]. Traffic visual analysis systems based on a virtual reality GIS represent the standard by which traffic data are controlled and developed. Aside from the fundamental GIS mutual functions, the system put forward also contains smart functions for visual analysis and forecast accuracy.

There is also a study that addresses the concept of smart and connected communities (SCCs), which are no longer solely defined as smart cities [44], [51], [52]. Big data analytics in cyber-physical systems, which are engineered systems that are built from, and depend upon, the seamless integration of computational algorithms and physical components, will enable the move from the IoT to real-time control and toward the SCC [49]. SCCs were conceived to represent earlier requirements (e.g., protection and redevelopment) in a cooperative way, and the requirements for current living (habitability) and planning for the future (sustainability). In consequence, the final objective of SCCs is to improve habitability, protection, redevelopment, and the attainability of a desirable society. This study uses mobile crowdsourcing and cyber-physical cloud computing for these two essential IoT technologies.

F. Challenges in Applications

There is also a common challenge in infrastructure-support applications in terms of efficiency. The more efficient the underlying infrastructure, the larger the number of facilities the next generation will support. For all domains (power prediction, user behavior, healthcare, content recommendation systems, and the smart city), a more efficient infrastructure is crucial in order to support efficient ML algorithms and to develop new ones. The models may scale efficiently with the amount of data represented in the big data ecosystem, as well as with the algorithms in charge of offering enhanced performance.

VII. Future Research Topics

We reviewed data types, analysis methods, data security, and applications related to network big data. This review shows that the data retrieval process is focused more and more on streaming and multiple sensor data. The analysis method mainly relied on a variant of MapReduce and ML. Data security is a potential problem in the era of big data.

The current research outcomes have indicated that data are no longer just data [57], [67], [68]. The value in the updates in big data lies in the data types, analysis algorithms, or new products. In the past few years, the growth in big data has been closely related to mobile and smart devices. The increasing popularity of the IoTs has also generated new types of big data, and various types of networking facilitate the interconnection of multivariate networking data. The relevant smart applications for big data have integrated media, communications, social networking, and sensors. The expectations for data collection are also getting critical, with only useful data being collected to solve urgent issues. With regard to the development of big data, current facilities have provided greater convenience and mobility, allowing more flexible and effective processes for terminal devices and material collection. The digitization of various types of information has led the circulation, exchange, processing, and application of the information toward more organized standards and structures. The application of data has become more direct and is moving into real time. The digitized trades have completely changed human trading behaviors and capital flows, enabling trading data to be maintained and applied to analyzing economic principles, as well as offering a reference for future business model designs. The interactions between humans and machines generate abundant big data, from which the potential can be extracted to design a fit mode for human lives. The cost of acquiring data has been lowered, benefiting real-time collection and processing of big data, changing the relationship between decision-making and information, and increasing the chances of extracting the right model from the data. Data technology has been widely accepted as an optimization tool, or is intended for complete innovation. Data collection, updates, recognition, and correlation will become more automatic. Since a lot of countries have started to adopt new data security technology and new data protection laws, supervision of big data security will be stricter. As for data security, the public is more concerned with protection of personal privacy, rather than trade secrets. Besides, governments enjoy having the most data (other than media and social media apps), with their data covering resources, finance, transportation, security, medical care, the environment, food, and so on. The open data policies of governments matter critically for the development of the entire data industry. All points where big data lands are linked with the industries, and those industries (fully influenced by the Internet), such as finance, medical care, and e-commerce, can easily be digitized. Big data has been gradually applied to seek solutions for each industry.

References

[1] C. Min, M. Shiwen, and L. Yunhao, “Big data: A survey,” *Mobile Netw. Appl.*, no. 2, pp. 171–209, Apr. 2014.

[2] S. Chris, U. Matzat, and U.-D. Reips, “Big data: Big gaps of knowledge in the field of internet science,” *Int. J. Internet Sci.*, vol. 7, no. 1, pp. 1–5, 2012.

[3] J. Minho, T. Maksymyuk, R. L. Batista, T. F. Maciel, A. L. F. de Almeida, and M. Klymash, “A survey of converging solutions for heterogeneous mobile networks,” *IEEE Wirel. Commun.*, vol. 21, no. 6, pp. 54–62, Dec. 2014.

[4] J. Minho, L. Han, N. D. Tan, and H. P. In, “A survey: Energy exhausting attacks in MAC protocols in WBANs,” *Telecommun. Syst.*, vol. 58, no. 2, pp. 153–164, 2015.

[5] B. Mitra, N. Meratnia, and P. J. M. Havinga, “On the use of mobility data for discovery and description of social ties,” in *Proc. 2013 IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2013, pp. 1229–1236.

[6] M. Marco and S. Valtolina, “Towards a user-friendly loading system for the analysis of big data in the internet of things,” in *Proc. 2014 IEEE 38th Int. Comput. Softw. Appl. Conf. Workshops*, 2014, pp. 312–317.

[7] S. H. Thiago, P. O. S. Vaz De Melo, J. M. Almeida, and A. A. F. Loureiro, “Large-scale study of city dynamics and urban social behavior using participatory sensing,” *IEEE Wirel. Commun*, vol. 21, no. 1, pp. 42–51, Feb. 2014.

[8] H. Guo, X. Li, W. Wang, Z. Lv, C. Wu, and W. Xu, “An event-driven dynamic updating method for 3D geo-databases,” *Geo-Spatial Inf. Sci.*, vol. 19, pp. 140–147, 2016.

[9] S. Tianyun, Z. Cao, Z. Lv, C. Liu, and X. Li, "Multi-dimensional visualization of large-scale marine hydrological environmental data," *Adv. Eng. Softw.*, vol. 95, pp. 7–15, 2016.

[10] J. Sunil and C. Lingam, "Reality mining based on social network analysis," in *Proc. 2015 Int. Conf. Commun., Inf. Comput. Technol.*, 2015, pp. 1–6.

[11] C. W. Boulila and I. R. Farah, "Improvement of satellite image classification: Approach based on Hadoop/MapReduce," in *Proc. 2016 2nd Int. Conf. Adv. Technol. Signal Image Process.*, 2016, pp. 31–34.

[12] M. Sarwat, "Interactive and scalable exploration of big spatial data—A data management perspective," in *Proc. 2015 16th IEEE Int. Conf. Mobile Data Manage.*, 2015, pp. 263–270.

[13] D. Xuejun, Y. Tian, and Y. Yu, "A real-time big data gathering algorithm based on indoor wireless sensor networks for risk analysis of industrial operations," *IEEE Trans. Ind. Informat.*, vol. 12, no. 3, pp. 1232–1242, Jun. 2016.

[14] R. M. Simmonds, P. Watson, J. Halliday, and P. Missier, "A platform for analysing stream and historic data with efficient and scalable design patterns," in *Proc. 2014 IEEE World Congr. Serv.*, 2014, pp. 174–181.

[15] B. Arian, A. Finamore, P. Casas, L. Golab, and M. Mellia, "Large-scale network traffic monitoring with DBStream, a system for rolling big data analysis," in *Proc. 2014 IEEE Int. Conf. Big Data*, 2014, pp. 165–170.

[16] A. Nishant and A. K. Sharma, "Proposed algorithms for effective real time stream analysis in big data," in *Proc. 2015 3rd Int. Conf. Image Inf. Process.*, 2015, pp. 348–352.

[17] J. Shatha, N. Dokoohaki, and M. Matskin, "OLLDA: A supervised and dynamic topic mining framework in twitter," in *Proc. 2015 IEEE Int. Conf. Data Mining Workshop*, 2015, pp. 151354–151359.

[18] Y. Xiaomeng, F. Liu, J. Liu, and H. Jin, "Building a network highway for big data: Architecture and challenges," *IEEE Netw.*, vol. 28, no. 4, pp. 5–13, Jul./Aug. 2014.

[19] H. Eszter, "Is bigger always better? Potential biases of big data derived from social network sites," *Ann. Amer. Acad. Political Social Sci.*, vol. 659, no. 1, pp. 63–76, 2015.

[20] W. Ke, X. Zhou, T. Li, D. Zhao, M. Lang, and I. Raicu, "Optimizing load balancing and data-locality with data-aware scheduling," in *Proc. 2014 IEEE Int. Conf. Big Data*, 2014, pp. 119–128.

[21] Z. Dongfang *et al.*, "Fusionfs: Toward supporting data-intensive scientific applications on extreme-scale high-performance computing systems," in *Proc. 2014 IEEE Int. Conf. Big Data*, 2014, pp. 61–70.

[22] L. Tonglin *et al.*, "ZHT: A light-weight reliable persistent dynamic scalable zero-hop distributed hash table," in *Proc. 2013 IEEE 27th Int. Symp. Parallel Distrib. Process.*, 2013, pp. 775–787.

[23] K. Jaein, N. Kim, B. Lee, J. Park, K. Seo, and H. Park, "RUBA: Real-time unstructured network big data framework," in *Proc. 2013 Int. Conf. ICT Convergence*, 2013, pp. 518–522.

[24] L. Rongxing, H. Zhu, X. Liu, J. K. Liu, and J. Shao, "Toward efficient and privacy-preserving computing in big data era," *IEEE Netw.*, vol. 28, no. 4, pp. 46–50, Jul./Aug. 2014.

[25] L. Chi, Z. Song, H. Song, Y. Zhou, Y. Wang, and G. Wu, "Differential privacy preserving in big data analytics for connected health," *J. Med. Syst.*, vol. 40, no. 4, pp. 1–9, 2016.

[26] S. Shan, "Big data classification: Problems and challenges in network intrusion prediction with machine learning," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 41, no. 4, pp. 70–73, 2014.

[27] R. Gentz, H. T. Wai, A. Scaglione, and A. Leshem, "Detection of data injection attacks in decentralized learning," in *Proc. 2015 49th Asilomar Conf. Signals, Syst. Comput.*, 2015, pp. 350–354.

[28] A. Saurabh and I. Chana, "A survey of clustering techniques for network big data," in *Proc. 2014 5th Int. Conf. Confluence Next Gener. Inf. Technol. Summit*, 2014, pp. 59–65.

[29] M. S. Ganesh and S. Ravi, "Network big data using apache hadoop," in *Proc. 2014 Int. Conf. IT Convergence Secur.*, 2014, pp. 1–4.

[30] X. Li and G. Cheng, "Research status and scientific thinking of big data," *Bull. Chin. Acad. Sci.*, vol. 6, pp. 647–657, 2012.

[31] Y. Tengfei, Y. Han, Y. Chen, and G. Chen, "WebVR—Web virtual reality engine based on P2P network," *J. Netw.*, vol. 6, no. 7, pp. 990–998, 2011.

[32] B. Elagib Sara, A. R. Najeeb, A. H. Hashim, and R. F. Olanrewaju, "Network big data solutions using mapreduce framework," in *Proc. 2014 Int. Conf. Comput. Commun. Eng.*, 2014, pp. 127–130.

[33] J. Wang, Y. Tang, M. Nguyen, and I. Altintas, "A scalable data science workflow approach for big data Bayesian network learning," in *Proc. 2014 IEEE/ACM Int. Symp. Big Data Comput.*, 2014, pp. 16–25.

[34] Z. Huang *et al.*, "Transient power quality assessment based on network big data," in *Proc. 2014 China Int. Conf. Electr. Distrib.*, 2014, pp. 1308–1312.

[35] S. Michał and A. Romanowski, "Sentiment analysis of Twitter data within big data distributed environment for stock prediction," in *Proc. 2015 Fed. Conf. Comput. Sci. Inf. Syst.*, 2015, pp. 1349–1354.

[36] A. Jalel, "The road to democracy: Modeling and analysis of an election big data," in *Proc. 2015 2nd Int. Conf. eDemocracy eGovernment*, 2015, pp. 14–15.

[37] Z. Kunpeng *et al.*, "A probabilistic graphical model for brand reputation assessment in social networks," in *Proc. 2013 IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2013, pp. 223–230.

[38] L. Zhihan, J. Chirivella, and P. Gagliardo, "Bigdata oriented multimedia mobile health applications," *J. Med. Syst.*, vol. 40, no. 5, pp. 1–10, 2016.

[39] L. Zhihan *et al.*, "Managing big city information based on WebVRGIS," *IEEE Access*, vol. 4, pp. 407–415, 2016.

[40] L. Xiaoming *et al.*, "WebVRGIS based traffic analysis and visualization system," *Adv. Eng. Softw.*, vol. 93, pp. 1–8, 2016.

[41] Y. Jiachen, H. Wang, Z. Lv, W. Wei, H. Song, M. Erol-Kantarci, and S. He, "Multimedia recommendation and transmission system based on cloud platform." *Future Gener. Comput. Syst.*, 2016.

[42] D. George and P. Demestichas, "Intelligent transportation systems," *IEEE Veh. Technol. Mag.*, vol. 5, no. 1, pp. 77–84, Mar. 2010.

[43] L. Tonglin *et al.*, "A convergence of key-value storage systems from clouds to supercomputers," *Concurrency Comput.: Pract. Experience*, vol. 28, no. 1, pp. 44–69, 2016.

[44] B. Wenjie, M. Cai, M. Liu, and G. Li, "A big data clustering algorithm for mitigating the risk of customer churn," *IEEE Trans. Ind. Informat.*, vol. 12, no. 3, pp. 1270–1281, Jun. 2016.

[45] Z. Yu, "Methodologies for cross-domain data fusion: An overview," *IEEE Trans. Big Data*, vol. 1, no. 1, pp. 16–34, Mar. 2015.

[46] E. P. Xing *et al.*, "Petuum: A new platform for distributed machine learning on big data," *IEEE Trans. Big Data*, vol. 1, no. 2, pp. 49–67, Jun. 2015.

[47] F. Wu *et al.*, "Weakly semi-supervised deep learning for multi-label image annotation," *IEEE Trans. Big Data*, vol. 1, no. 3, pp. 109–122, Sep. 2015.

[48] Y. Yang, F. Shen, H. T. Shen, H. Li, and X. Li, "Robust discrete spectral hashing for large-scale image semantic indexing," *IEEE Trans. Big Data*, vol. 1, no. 4, pp. 162–171, Dec. 2015.

[49] S. Jiang, X. Qian, T. Mei, and Y. Fu, "Personalized travel sequence recommendation on multi-source big social media," *IEEE Trans. Big Data*, vol. 2, no. 1, pp. 43–56, Mar. 2016.

[50] H. Song, D. Rawat, S. Jeschke, and C. Brecher, *Cyber-Physical Systems: Foundations, Principles and Applications*. Boston, MA, USA: Academic, 2016, ISBN 978-0-12-803801-7.

[51] Y. Sun, H. Song, A. J. Jara, and R. Bie, "Internet of things and big data analytics for smart and connected communities," *IEEE Access*, vol. 4, pp. 766–773, Mar. 2016.

[52] H. Song, S. Ravi, T. Sookoor, and S. Jeschke, *Smart Cities: Foundations, Principles and Applications*. Hoboken, NJ, USA: Wiley, 2017, ISBN: 978-1119226390.

[53] H. Song, G. Fink, and S. Jeschke, *Security and Privacy in Cyber-Physical Systems: Foundations and Applications*. England, U.K.: Wiley, 2017.

[54] Y. Zhang, L. Sun, H. Song, and X. Cao, "Ubiquitous WSN for healthcare: Recent advances and future prospects," *IEEE Internet Things J.*, vol. 1, no. 4, pp. 311–318, Aug. 2014.

[55] L. A. Tawalbeh, W. Bakheder, and H. Song, "A mobile cloud computing model using the cloudlet scheme for big data applications," in *Proc. 2016 IEEE 1st Int. Conf. Connected Health: Appl., Syst. Eng. Technol.*, 2016, pp. 73–77.

[56] S. Lohr, "The age of big data," *NY Times*, vol. 11, p. SR1, 2012.

[57] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of Google flu: Traps in big data analysis," *Science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[58] A. Menon, "Big data@ facebook," in *Proc. 2012 Workshop Manage. Big Data Syst.*, Sep. 2012, pp. 31–32.

[59] C. P. Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Inf. Sci.*, vol. 275, pp. 314–347, 2014.

[60] C. Snijders, U. Matzat, and U. D. Reips, "Big data: Big gaps of knowledge in the field of internet science," *Int. J. Internet Sci.*, vol. 7, no. 1, pp. 1–5, 2012.

[61] V. Marx, "Biology: The big challenges of big data," *Nature*, vol. 498, no. 7453, pp. 255–260, 2013.

[62] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Inf. Sci. Syst.*, vol. 2, 2014, Art.ID. 3.

[63] H. V. Jagadish *et al.*, “Big data and its technical challenges,” *Commun. ACM*, vol. 57, no. 7, pp. 86–94, 2014.

[64] H. Hu, Y. Wen, T. S. Chua, and X. Li, “Toward scalable systems for big data analytics: A technology tutorial,” *IEEE Access*, vol. 2, pp. 652–687, 2014.

[65] R. Kitchin, “The real-time city? Big data and smart urbanism,” *Geo J.*, vol. 79, no. 1, pp. 1–14, 2014.

[66] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, “Trends in big data analytics,” *J. Parallel Distrib. Comput.*, vol. 74, no. 7, pp. 2561–2573, 2014.

[67] J. Fan, F. Han, and H. Liu, “Challenges of big data analysis,” *Nat. Sci. Rev.*, vol. 1, no. 2, pp. 293–314, 2014.

[68] T. Boellstorff, *Coming of Age in Second Life: An Anthropologist Explores the Virtually Human*. Princeton, NJ, USA: Princeton Univ. Press, 2015.

[69] L. Aniello, R. Baldoni, and L. Querzoni, “Adaptive online scheduling in storm,” in *Proc. 7th ACM Int. Conf. Distrib. Event-Based Syst.*, Jun./Jul. 2013, pp. 207–218. doi: 10.1145/2488222.2488267.

[70] P. Basanta-Val, N. Fernández García, A. J. Wellings, and N. C. Audsley, “Improving the predictability of distributed stream processors,” *Future Gener. Comput. Syst.*, vol. 52, pp. 22–36, 2015.

[71] P. Basanta-Val and M. García-Valls, “A distributed real-time java-centric architecture for industrial systems,” *IEEE Trans. Ind. Informat.*, vol. 10, no. 1, pp. 27–34, Feb. 2014.

[72] M. Congosto, D. Fuentes-Lorenzo, and L. Sánchez, “Microbloggers as sensors for public transport breakdowns,” *IEEE Internet Comput.*, vol. 19, no. 6, pp. 18–25, Nov./Dec. 2015.

[73] P. Basanta-Val, N. C. Audsley, A. Wellings, I. Gray, and N. Fernandez-Garcia, “Architecting time-critical big-data systems,” *IEEE Trans. Big Data*, vol. 2, no. 4, pp. 310–324, Dec. 2016. doi: 10.1109/TBDATA.2016.2622719.

[74] T. Higuera-Toledano, “Java technologies for cyber-physical systems,” *IEEE Trans. Ind. Informat.*, 2016.