REVIEW ARTICLE

Big spatial vector data management: a review

Xiaochuang Yao and Guoqing Li

Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China

ABSTRACT
Spatial vector data with high-precision and wide-coverage has exploded globally, such as land cover, social media, and other datasets, which provides a good opportunity to enhance the national macroscopic decision-making, social supervision, public services, and emergency capabilities. Simultaneously, it also brings great challenges in management technology for big spatial vector data (BSVD). In recent years, a large number of new concepts, parallel algorithms, processing tools, platforms, and applications have been proposed and developed to improve the value of BSVD from both academia and industry. To better understand BSVD and take advantage of its value effectively, this paper presents a review that surveys recent studies and research work in the data management field for BSVD. In this paper, we discuss and itemize this topic from three aspects according to different information technical levels of big spatial vector data management. It aims to help interested readers to learn about the latest research advances and choose the most suitable big data technologies and approaches depending on their system architectures. To support them more fully, firstly, we identify new concepts and ideas from numerous scholars about geographic information system to focus on BSVD scope in the big data era. Then, we conclude systematically not only the most recent published literatures but also a global view of main spatial technologies of BSVD, including data storage and organization, spatial index, processing methods, and spatial analysis. Finally, based on the above commentary and related work, several opportunities and challenges are listed as the future research interests and directions for reference.

1. Introduction
Over the last 10 years, “Big Data” has become a state of the art technical term in academia, industry, business, as well as politics. Big data refers to the rapid growth of data that is beyond the ability of traditional databases and software tools in terms of data acquisition, storage, processing, and usage (Chen et al., 2013; Manyika et al., 2011). Tagged with the “3Vs+” characteristic, big data has established new opportunities (Stantic & Pokorný, 2014) and significant challenges (Chen et al., 2013; Kanchi, Sandilya, Ramkrishna, Manjrekar, & Vhadgar, 2015).
in big data management, which force the revolutionary measures needed urgently (Chen et al., 2013; Mcaffee & Brynjolfsson, 2012).

Big spatial data is probably one of the most active areas in both amount of data growth and technical applications (Lee & Kang, 2015). With the development and improvement of a new generation of high-performance computing technologies, such as cloud computing, NoSQL databases, and others, the relevant theories and methods have begun to penetrate into the field of geographic information science (GIS) gradually (Yang, Huang, Li, Liu, & Hu, 2016; Yang, Yu, Hu, Jiang, & Li, 2017). Especially for the remote sensing (RS) big data, so far, a systematic platform for collecting, storing, organizing, analyzing, visualizing, and applying RS big data based on cloud storage and high-performance computing has been initially formed (Ma et al., 2015; Mulyono & Fanany, 2015; Pekturk & Unal, 2017; Sun, Chen, Chi, & Zhu, 2015; Wang, Ma, Yan, Chang, & Zomaya, 2018; Wang et al., 2015). However, studies on big spatial vector data (BSVD) are relatively fewer because the vector data-sets often involve a number of special factors, such as national economic, national defense, and other infrastructure constructions, which have led to difficulties in large-scale data sharing and acquisition.

Big spatial vector data (BSVD) is a breadth of spatial data types, which can be represented by points, lines, or polygons (areas) (Shekhar, Evans, Gunortic, Yang, & Cugler, 2014; Tong, Ben, Liu, & Zhang, 2013). Generally speaking, BSVD can include all of the surveying and mapping data, location-based data, social media data, as well as Internet of Things data. Data management is also a more complex and broad domain covering data storage, index and query, data processing, and analysis (Siddiqa et al., 2016). The existing work in BSVD management has mostly emphasized on some characteristics (volume, variety, or velocity) of big spatial data, and solved certain problems in the technical level or applications. Although there already have several studies related to big spatial data (Eldawy & Mokbel, 2015a; Lee & Kang, 2015; Yang, Yan, & Nebert, 2013) and big data management (Kanchi et al., 2015; Karim et al., 2017; Li, Xu, Jiang, & Zhang, 2014; Siddiqa et al., 2016; Storey & Song, 2017), however, by now, no scholar has directly focused on big spatial vector data (BSVD) management and reviewed the latest literatures to provide a comprehensive survey of the existing techniques and technologies.

This review paper mainly focuses on spatial vector data management in the era of big data. The big spatial data, data storage and organization, data processing, and spatial analysis are discussed, respectively. In the context of BSVD management, this study categories the existing techniques and technologies, as well as highlighting the mainstream academic views to help the readers to better understand and handle the problems from BSVD management efficiently. In addition, this review summarizes the characteristics and domains of big spatial data, and also overviews the BSVD management. Moreover, a broad of literatures on the vector data model, data storage, spatial index, pre-processing, spatial query, visualization, and spatial analysis of BSVD are provided and classified. Finally, future research interests and directions are contributed as a guide for researchers.

The rest of this paper is organized as follows: Section 2 presents a general overview of big spatial data and BSVD management. Sections 3 and 4 discuss the techniques and technologies for data storage (data model, storage, and index) and processing (pre-processing, spatial query, visualization, and spatial analysis), respectively. Section 5 highlights the future interests and directions for big spatial vector data management and conclusion of this study is provided in Section 6.
2. Big spatial data

2.1. Big data in GIS

GIS is a comprehensive technology (Zhao, Chen, Ranjan, Choo, & He, 2015), which involves the geography, mapping science, computer science technology, and other disciplines. With the development of computer science technology and the evolution of computing model, GIS architecture and application mode are changed constantly. From desktop GIS (1960s) to the Web GIS (1980s), and the distributed GIS (1990s), to the cloud GIS (2010s), it is well known that the development of GIS is greatly influenced by computer science technology (Yang, Raskin, Goodchild, & Gahegan, 2010), and also lags behind it simultaneously. Needless to say, driven by the wave of big data, GIS has also entered a new era, big data GIS (Li & Li, 2014), and a large number of scholars have already put forward some new concepts and new ideas in the era of big data. As the Table 1 shows, it lists new thinking about GIS in the era of big data.

In addition, some scholars have called for the core of geography to be carried out during the shift from “data big” to “big data”. In the era of big data, through the use of cloud computing and Internet of Things technologies, we should not only build the next generation GIS, but also need to emphasize the basic theory of spatial analysis, modeling, and optimization, and adhere to GIS philosophy itself and spatial thinking (Zhang, 2014). Through the comparative studies of new ideas and new concepts, we argue that: the big data age already has a certain influence on GIS, but the current GIS is still in the stage of “data big”, and the core proposition of GIS and the basic paradigm have not been changed fundamentally because of the influence of big data. Really making the spatial data play its proper value, it requires a large number of scholars to continue the work, and this paper hopes to facilitate the future researchers.

2.2. Big spatial data

Data are often referred to as the “blood” of GIS. This point is consistent with big data, which is based on “data” and driven by “data”. Big data not only reflects large data volume, but now its characteristics have increased from “3Vs (Volume, Velocity, Variety)” to “4Vs (+ Veracity)”, even to “5Vs (+ Value)” (Li & Li, 2014). Massive data have been the hot topic in GIS for several years. According to the characteristics and sources of spatial data, as the Figure 1 shows, we summarize the GIS data as five categories, namely remote sensing data, surveying and mapping large data, location-based data, social network data, and Internet of Things data. Usually, there is a certain intersection between these categories.

(1) Remote sensing data (Chi et al., 2016; Mulyono & Fanany, 2015): the increase of remote sensing data is from quantity to quality. There are more and more satellite platforms, including aerospace, aviation as well as the near space. The space, time, and spectral resolution are constantly improved, which can be obtained with the TB level, and stored with PB.

(2) Surveying and mapping data (Huang et al., 2016; Lu, Yuan, & Yu, 2017; Wang, Guan, & Wu, 2017): it generally includes geographical situation, industry geographic, thematic mapping data, such as 4D (DLG, DRG, DOM, DEM) digital products, land use, and other national basic surveying and mapping data. In recent years, with the
| New concepts and ideas                     | Keywords or highlights                                                                                                                                                                                                 | Contributors                        |
|-------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------|
| Big geo data                              | The quality of big (geo) data and three methods for quality improvement and assurance                                                                                                                                  | Goodchild (2013)                    |
| Big data GIS                              | Big data GIS is from offline analysis/model-driven/simple visualization to dynamic/data-driven/visual analysis in real time                                                                                                                                                  | Li and Li (2014)                    |
| Generalized GIS                           | The characteristics of generalized GIS and technical challenges from data collection and cleaning, management and integration, data analysis, and computing                                                                 | Lu and Zhang (2014)                 |
| Open GIS                                  | Eight dimensions of open GIS in big data era                                                                                                                                                                                                                                      | Sui, Ye, and Gan (2014)             |
| Big spatial data/Spatial big data         | Big data technologies for systems; traditional/emerging spatial big data with different data types; metrics for spatial big data benchmarks                                                                                                                                   | Shekhar et al. (2014), Eldawy and Mokbel (2015a) |
| Urban computing                           | A process of acquisition, integration, and analysis of big and heterogeneous data generated by a diversity of sources in urban spaces to tackle the major issues that cities face                                                                                      | Zheng (2015, 2017)                  |
| Social sensing                            | A big geo-data-based approach to understand our socioeconomic environments                                                                                                                                                                                                     | Liu et al. (2015)                   |
| Data-driven geography                     | From a data-scarce to a data-rich environment and use big data to inform both geographic knowledge-discovery and spatial modeling                                                                                                                                            | Miller and Goodchild (2015)         |
| Geospatial big data                       | Geospatial big data refers to spatial data-sets exceeding capacity of current computing systems                                                                                                                                                                              | Lee and Kang (2015)                 |
| New definition for GIS                    | A multiple discipline science and technology, integrating data acquisition, extraction, management, knowledge discovering, spatial sensing and recognition, as well as intelligent location-based services of any physical objects and human activities around the earth and its environment | Li (2016)                           |
| Big earth data                            | Big Earth Data from space can be defined as big data obtained via the methodologies of Earth observation, such as data from space borne, airborne, and ground sensors                                                                                      | Guo, Wang, and Liang (2016), Guo (2017) |
| Spatio-temporal big data/Big spatio-temporal data | Beside the huge achieved volume of the data, space and time are two fundamental characteristics that raise the demand for processing spatio-temporal data                                                                                                          | Alarabi and Mokbel (2017), Wang (2017) |
| Remote sensing big data                   | Data infrastructure for remote sensing big data: integration, management and on-demand service. The RS “Big Data” not merely refers to the volume and velocity of data that outstrip the storage and computing capacity, but also the variety and complexity of the RS data | Li and Huang (2017), Ma et al. (2015) |
development of new surveying and mapping equipment and technology, the era of big data has been accelerated, such as point cloud, mobile mapping, which can quickly and efficiently obtain the spatial distribution data of the measurement area.

(3) Location-based data (Liu, Fang, Guo, & Gao, 2014; Liu et al., 2015; Zhuang et al., 2017): the geographic and human social information data that usually contain spatial location and time identification are known as location-based data. Location-based data are mainly with GPS, BDS, and other positioning systems generated by the smart phones, the collection data in the field, the traffic trajectory data, etc. Big location-based data have become an important strategic resource that is used to sense the activities of human social groups.

(4) Social media data (Cervone et al., 2016; Magdy, Mokbel, Elnikety, Nath, & He, 2016; Tsou, 2015): internet data with spatial location, including the user’s web page, data in social media, such as WeChat, Facebook, Twitter, and other social software. At present, social media data have played an important role in online public opinion, natural disaster monitoring, and environmental law enforcement.

(5) Internet of Things data (Alelaiwi, 2017; Ding, Chen, & Yang, 2014): all kinds of sensor monitoring data, including environmental protection, meteorology, water, pipelines monitoring, and wearable devices, intelligent household and others. Compared with the traditional internet, the Internet of Things data are generated with higher frequency and greater variety.

2.3. BSVD management

Data management is a multidisciplinary, which is using computer hardware and software technology to collect, store, process, and apply data effectively (Kanchi et al., 2015). Its aim

Figure 1. The classification diagram of big spatial data.
is to take advantage of the data to make business intelligence and scientific decisions (Siddiqa et al., 2016). In the era of big data, change of data management is not just data itself, but also the computer hardware and software technology. Big data management is blessed with new challenges in terms of data storage, processing, and governance (Mcafee & Brynjolfsson, 2012). Figure 2 depicts the key process flow of the big spatial vector data (BSVD) management. And it needs to reconsider the current technology environment for every step, including storage, spatial index, processing, and applications.

Big data management has been successfully used in various industries, such as information security (Xu, Jiang, Wang, Yuan, & Ren, 2014), education (Zhang, 2017), health (Bradley, 2013), archeology (McCoy, 2017), and so on. For the BSVD management, the Table 2 lists the existing frameworks or systems, which are compared from three aspects, namely architecture, spatial index, and spatial query.

3. Data storage and organization

The data storage and organization is the basic step of data management. Existing storage strategies exist in various ways. In order to provide a better data acquisition interface in big spatial vector data system, the data model, storage mode, and spatial index are three aspects that must be considered. The data model is designed to adapt to the storage mode, and spatial index will speed up the efficiency of big data retrieval.

3.1. Vector data model

The spatial elements are represented by points, lines, and polygons in the vector data model (Shekhar et al., 2014). Compared with the raster data model, it has the advantages of high precision, small volume, and good quality. According to the storage relationship between spatial and attribute data, the vector data model can be classified into a geographical relation
model and an object-oriented model. The object-oriented model has been very popular because of the advantages (Wojda & Brouyère, 2013) of being easy to understand, expand, represent, and realize. Although there are some common vector data structures provided by OGC standards (Zheng & Fu, 2013), such as KML, GML, WKT/WKB, and GeoJSON, not all of them are suitable for big data storage structures. As the Figure 3 shows, new spatial data model should be considered from both the vector data model and the data model in database.

The key–value model is now the mainstream of the storage model in a large number of NoSQL databases. In the key–value model, each record consists of two parts, also known as “Key/Value Pair”, which supports simple data operation. Based on the simple key–value

---

**Table 2. Existing frameworks or systems for big spatial vector data.**

| Name            | Architecture | Spatial index | Spatial query            | Contributors                   |
|-----------------|--------------|---------------|--------------------------|--------------------------------|
| MD-HBase        | HBase        | Quad Tree; K-d tree | Range Query; KNN         | Nishimura et al. (2011)         |
| CloST           | Hadoop       | R-tree        | Range Query; Spatial Join; KNN | Tan, Luo, and Ni (2012)         |
| Hadoop-GiS      | Hadoop       | Grid          | Range Query; Spatial Join; KNN | Aji et al. (2013)               |
| VegaSTDE        | Hadoop HBase | Quad Tree     | ST-Range Query           | Zhong, Fang, and Zhao (2013)    |
| SpatialHadoop   | Hadoop       | R-tree; Quad-tree; Grid; K-d tree | Range Query; Spatial Join; KNN | Eldawy and Mokbel (2013)        |
| Parallel SECONDO | Parallel DB  | –             | Range Query; Spatial Join | Lu and Guting (2014)            |
| ScalaGiST       | Hadoop       | B+–tree; R-tree | Range Query; KNN         | Lu et al. (2014)                |
| GeoMesa         | Hbase        | GeoHash       | Range Query              | Hughes et al. (2015)            |
| GeoSpark        | Spark        | R-tree; Quad-tree | Range Query; Spatial Join; KNN | Yu, Wu, and Sarwat (2015)       |
| Sphinx          | Parallel DB  | R-tree; Quad tree | Range Query; Spatial Join | Eldawy, Elganainy, et al. (2015) |
| Simba           | Spark        | R-tree        | Range Query; Spatial Join; KNN | Xie et al. (2016)               |
| ST-Hadoop       | Hadoop       | ST-index      | ST-range query; ST-join   | Alarabi and Mokbel (2017)       |
| LandQ®          | Parallel DB  | Grid          | Range Query              | Yao, Yang, et al. (2017)        |

**Figure 3.** Big spatial vector data model.
model, spatial vector data can be imported and stored in a big data platform (Wang, Chen, & Liu, 2013; Zheng & Fu, 2013), such as HDFS and HBase, with new spatial vector data models. As the Figure 4 shows, the data structure of the GeoCSV stores spatial geometric elements in the cloud environment based on the object-based vector data model. GeoCSV uses the simple key–value storage model and the OGC-WKT format to describe the spatial geometry. It takes the advantage of CSV (Comma-Separated Values) files to organize the spatial vector data, that is, each record represents only one spatial geometry object. This is in line with the cloud computing platform, which is conducive to spatial data segmentation, processing, and analysis. Based on GeoCSV, it has several advantages with parallel computation, network transmission, and expansion.

3.2. Data storage

Spatial database is an effective means to manage vector data, which is the basis of vector data query, analysis, and application. For nearly half a century, the spatial database management technology has mainly experienced four stages of evolution, namely the file system (1970s), the file–relationship hybrid system (1980s), the spatial database engine (1990s), and the space of the object-oriented relational database management system (twenty-first century). In the era of big data, with the development of computer science technology, new research results appear in spatial database. At present, the storage and organization methods of large vector data are mainly used in the following three modes, namely the relational database, NoSQL database, and distributed file system.

3.2.1. Relational database

Based on relational databases, such as Oracle, PostgreSQL, distributed storage of spatial vector data has been a hot topic on distributed spatial database, and a long-term work focused on the design and implementation of a distributed spatial database engine (SDE). In general, the distributed SDE supports secondary application development, which is the

![Figure 4. The structure of big spatial vector data, GeoCSV.](image-url)
bridge between users and databases. Niharika (Ray, Simion, Brown, & Johnson, 2013) is a distributed spatial database schema based on PostgreSQL and PostGIS, which performs spatial data partitioning, reading and writing operations in the cloud environment. Sphinx (Eldawy, Elganainy, Bakeer, Abdelmotaleb, & Mokbel, 2015) extends the Apache Impala to support the SQL language for spatial queries. Oracle RAC (Real Application Clusters) and Oracle Spatial provide distributed spatial vector data management systems for parallel query (Hameurlain & Morvan, 2016). Based on the SQLServer database and ArcGIS server, LandQv1 (Yao et al., 2017) is built to manage the arable land quality big data. The advantage of a vector data management system based on relational databases is the small amount of data and database migration, at the same time, in the service and application layer, it will have efficient integration with the original systems.

3.2.2. NoSQL database

NoSQL, or Not Only SQL, refers to non-relational databases. Currently, mainstream NoSQL databases include column storage and key-value modes, such as MongoDB, BigTable, Hbase, and Redis … Due to the advantages of distribution, scalability, and no predefined table structure, in recent years, NoSQL databases have been favored by researchers and commercial companies, also in the big spatial vector data management field. Based on Hbase, MD-hbase system (Nishimura, Das, Agrawal, & El Abbadi, 2011) is used to manage LBS data, and spatial partition based on K-d index tree is constructed to support the range and adjacent query functions. The GeoMesa (Hughes et al., 2015) is launched by the federal computer research center and with an HBase database, the GeoHash index-based spatial range query function is implemented. Developed by the Chinese Academy of Sciences (CAS), VegaSTDE platform (Zhong, Fang, & Zhao, 2013), including the VegaStore storage layer, is designed and implemented, respectively, with a mixed storage structure of HDFS and Hbase database to support an improved quad-tree index and spatio-temporal data query functions. Based on the NoSQL database, storage of big spatial vector data (BSVD) is not only convenient to deploy, but can also effectively integrate multi-source spatial data, which is beneficial to parallel computing.

3.2.3. Distributed file system

A distributed file system is connected to the nodes through a computer network to manage the storage resources. It has good support for ordinary computers and also has a good scalability and fault tolerance in clusters. Among them, the Hadoop Distributed File System (HDFS) is a typical distributed file system. For spatial vector data storage based on HDFS, CloST (Tan, Luo, & Ni, 2012) is built to store the GPS data with an innovative data model (ID/location/time), and supports a parallel implementation of the R-tree index. Hadoop-GIS (Aji et al., 2013) is developed to establish a set of high-performance spatial data warehouse system. Through the spatial partition to achieve a variety of spatial data query, the results show that the system is superior to the parallel spatial data relational database system. An open source framework, SpatialHadoop (Eldawy & Mokbel, 2013, 2015b) covers the high-level programming language based on the Pigeon (Eldawy & Mokbel, 2014), spatial data index, spatial query, and visualization (Eldawy, Mokbel, & Jonathan, 2016), application (Eldawy et al., 2015) to solve the basic problems of big spatial vector data management, which has a significant representative and reference value in industry. GeoSpark (Yu, Wu, & Sarwat, 2015) based on HDFS is designed with a three-layer architecture, namely, the Apache Spark layer, spatial
data distribution layer, and spatial query operational layer. Simba (Xie et al., 2016) uses a hybrid architecture based on HDFS and RDBMSS to store the vector data, while the Spark SQL engine is extended to support data query and analysis functions.

3.3. Spatial index

Spatial index can be simply understood as a data organization structure that can quickly and randomly access individual or multiple spatial objects within it. Today, massive data storage is no longer a file or a machine. Distributed storage has become a mainstream solution, such as Google GFS, Hadoop HDFS, etc. A distributed spatial index is established on a distributed storage system to meet the rapid retrieval of huge amounts of spatial data. As a result, the advantages and disadvantages of the spatial index are both limited by the distributed storage system. Spatial data are stored in different nodes in a cluster, and there is a certain communication protocol between name node and data nodes. As the Figure 5 shows, the distributed index for the big spatial vector data is mainly including the local index, global index, and mixed index of both.

In the distributed spatial index, the data partition takes on a very important role. Spatial data partition refers to the process of dividing a spatial data set into several data blocks according to a certain demarcation rule. Traditional attribute data partition methods, such as ID division or random partition, are not ideal for dividing spatial data (Yao et al., 2017). For the spatial data, a good spatial data partition strategy should ensure both optimal performance of spatial operation and data balance in the cluster (Wei et al., 2015). The spatial partition methods can be summarized into three categories (Eldawy, Alarabi, & Mokbel, 2015; Yao et al., 2017), namely space partition, data partition, and spatial filling curve partition. Based on the above partition methods, the corresponding spatial indexes are built for big spatial vector data in clusters, such as the k-d tree (Wei et al., 2015), Grid, G-tree (Zhong, Li, Tan, Zhou, & Gong, 2015), HQ-tree (Feng, Tang, Wei, & Xu, 2014), and others (Al-Badarneh & Al-Alaj, 2011; Li & Zheng, 2013; Scitovski & Scitovski, 2013; Whitman, Park, Ambrose, & Hoel, 2014).

Figure 5. Distributed spatial index for big spatial vector data.
4. Data processing and analysis

This section discusses the processing and analysis methods in the big spatial vector data (BSVD) management. Data processing is a wide range of domains covering the whole data flow from pre-processing to data application. The geometric computing algorithms are always very complex and time-consuming, which makes big spatial data processing very slow, even impossible (Ray et al., 2013). This paper just touches upon data processing and analysis methods influenced by big data technologies. As the Figure 6 shows, four taxonomies of BSVD processing and analysis are proposed in recent literatures, namely pre-processing, spatial query, visualization, and spatial analysis.

4.1. Data pre-processing

Data pre-processing is becoming more and more complicated because of explosive growth of data volume. However, it is still a very critical step before to be used deeply, especially for data quality (Taleb, Dssouli, & Serhani, 2015). For the big spatial vector data, the pre-processing is involved in data cleansing, coordinate/projection transformation, as well as data format conversion. These processing methods are mostly object-oriented, therefore, they are very suitable for parallelization.

Generally, the data cleansing work will be done before the data enters the application system. For the spatial vector data, the corrupt or inaccurate records will be detected and corrected (or removed) from the whole data-set. The transformation between geographic coordinates and plane coordinates is an approach used commonly. Focusing on cylindrical projection, a linear rule approximation model (LRA-model) (Ye et al., 2016) is presented and computed based on spatial grids. This model constructs linear polynomials to approximate

![Figure 6. Data processing and analysis of BSVD.](image-url)
the transformation rule and gets good results. A parallel map projection framework (Tang & Feng, 2017) for big spatial vector data (BSVD) is proposed with a layered architecture that couples capabilities of cloud computing and high-performance computing accelerated by graphics processing units (GPUs). Due to the different acquisition means, there are many kinds of data formats to organize spatial vector data, such as Shapefile, KML/KMZ, DXF/DWG, GPX, and others. In spite of OGC having the corresponding documentation standards and interoperability protocols, however, because of the function limitation of processing tools (Shen, Wong, Camelli, & Liu, 2013), it is very difficult to implement the format conversion for large data-sets. Data conversion is the normalization and aggregation from one format to another (Jhummarwala, Mazin, & Potdar, 2016). For the arable land quality management in Hadoop, we develop a tool to convert the data from Shapefile to GeoCSV with MapReduce program. ESRI GIS Tools for Hadoop (ArcGIS, 2017) supports data format conversion from Shapefile to GeoJSON.

4.2. Spatial query

In GIS applications, spatial query operations are the basis of spatial analysis and are also the window of the GIS application system for users. Spatial query is always based on spatial index mechanism to find the spatial data of this condition from the database. Therefore, the efficiency of execution relies heavily on the performance of the spatial index algorithm (Zhong et al., 2012).

The common query operations of spatial vector data include the range query, spatial join, and k-Nearest Neighbor (KNN). As shown in the third column of Table 2, based on different platforms, such as GPU (Zhang & You, 2012) and Hadoop (Bellur, 2014), these query methods are implemented in different frameworks or systems. Additionally, some professional query engines, such as TOUCH (Nobari et al., 2013), Phidj (Fries, Boden, Stepiein, & Seidl, 2014), and AQWA (Aly et al., 2015), are developed for the large-scale spatial query processing in cloud or distributed systems (Bellur, 2014; You, Zhang, & Gruenwald, 2015a, 2015b).

4.3. Visualization

The visualization of spatial data has become a major analysis method in the era of big data. For visualizing the large spatial data-set, OGC Web Map Service (WMS) has provided a simple HTTP interface for requesting geo-registered map images (returned as JPEG, PNG, etc.) from one or more distributed geospatial databases (Yao, Zhu, Yun, Peng, & Li, 2017). Through the caching of map tile pyramids, the WMS can make the raster or big spatial vector data display and zoom in or out on-the-fly in a client browser in WebGIS applications (Yao et al., 2017). However, the existing solutions to large-scale map data are not ideal (Guo et al., 2015; Lin, Zhou, & Xia, 2016). On the one hand, the map slicing process takes a long time or cannot even be completed, on the other hand, the amount of spatial data executable at one time is limited for tiling, resulting in too many services and paths to manage and load in the client. In order to meet the rapid demand for visualization of large-scale vector data, the algorithm is implemented in parallel. HadoopViz (Eldawy et al., 2016) implements the map tile pyramid model with MapReduce by employing a three phase technique; partitioning–plotting–merging for tile and heat map, and it also provides an extensible interface to add new visualization types by users (Eldawy & Mokbel, 2015b). Moreover, how to manage the billions of map tiles
generated by the slice tool is also a significant problem. Based on the NoSQL database, a parallel storage and management of map tiles is developed to speed up the time of map visualization (Lin et al., 2016).

As the Figure 7 shows, the tile pyramid model for map visualization is not just based on images (Li, Hu, Zhu, Li, & Zhang, 2017). In recent years, the vector tile technology has been a new approach to deliver visuals of large vector data-sets with flexible styles and rich tight-binding attribute information. With map vector tiles, there is no need to head back to the server and fetch a different set of tiles if you want to filter the output or change the style of the geometry features in the client (Yu et al., 2017). So far, commercial software and open sources (Mapbox, 2018) are emerging in supporting the vector tiles, including ESRI products, MapBox, and others.

4.4. Spatial analysis

Spatial analysis is different from the traditional statistics (Zhang, 2014) and it is characteristic with location, direction, and scale, etc. Spatial analysis in the era of big data is more complex, as well as more valuable and useful. This paper focuses on computational-centric analysis methods, such as spatial overlay, spatial clustering, spatial interpolation, buffer, clipping, and Voronoi analysis, which do not involve application-oriented analysis methods.

Spatial overlay analysis is a time-consuming task and complex geometric algorithm. Parallelized algorithms for spatial overlay processing are required (Zhu, Huo, & Qiu, 2015). A parallel point-polygon overlay analysis (Zhou et al., 2015) is implemented on a Linux-based cluster system using the MPI cluster-computing and OpenMP multi-core paralleling computing tool. Based on the MapReduce program, a parallel method is implemented with a grid index for GIS polygon overlay processing (Wang, Liu, Liao, & Li, 2015). A double-index and data divide-and-conquer-based parallel point-polygon overlay method (Zhou et al., 2015) is proposed in a distributed computing environment to speed up the executing efficiency for big spatial vector data. Spatial clustering is a primary data mining method for knowledge discovery in spatial database. A density-based algorithm (MDBSCAN) (Schoier & Borruso, 2015) is used to discover the clusters of units in large spatial sets.

![Image or Vector?](Figure 7. Tile pyramid model for map visualization.)
zoning can be seen using density-based clustering in the case of big data (Scitovski, 2018). Spatial interpolation (Yao, Zhu, Ye, Zhang, & Li, 2014) is adopted to compute the spatial distribution of unknown points from a sampling data-set. A parallelized Kriging interpolation (Wei et al., 2015) of big spatial data is developed to speed up the execution time of consequent interpolation. Based on CPU–GPU, a hybrid parallel spatial interpolation algorithm (Wang et al., 2017) is implemented for massive LiDAR point clouds. Buffer analysis and spatial clipping method are the basic and common spatial operations. The parallel algorithms for buffering (Fan, Ji, Gu, & Sun, 2014) and clipping (Puri & Prasad, 2014, 2015) are usually adopted to improve computational efficiency in large-scale data. The Voronoi analysis is a spatial proximity analysis method. Parallel computing for constructing Voronoi diagram (Bolcheva & Lévy, 2017; Starinshak, Owen, & Johnson, 2014) is to speed up data analysis as well as visualization. In addition to the above spatial analysis algorithms, a CG_Hadoop framework (Eldawy, Li, Mokbel, & Janardan, 2013) presents a set of fundamental computational geometry operators, namely, polygon union, skyline, convex hull, farthest pair, and closest pair for other geometric algorithms.

5. Future interests and directions

Through the above discussion about big spatial vector data (BSVD) management, many aspects have been achieved in this field. However, a large number of significant problems and challenges are still existed and need to be solved in the future. Based on our understanding and opinion, three potential research directions of BSVD management are proposed in this section, and it is hoped that they can inspire the interested researchers.

5.1. Spatio-temporal data model

The advancement of existing distributed storage systems, such as relational database, NoSQL databases, and HDFS, can accommodate large data-sets, but more emphasis is required to better support spatio-temporal data, which is a very important feature of big spatial vector data (BSVD) (Yang et al., 2013). Therefore, a spatio-temporal data model is needed to cover all of big data and also better support the upper spatial index and analysis in the cloud environment (Chen et al., 2015).

5.2. Visualization analysis

In the era of big data, the visualization of spatial data has become a significant analysis method, which is the most direct and effective means for understanding data. Visualization analysis is not a simple essential information display of big data, but driving complex analyses (Keim, Qu, & Ma, 2013). Especially, for the new emerging big spatial vector data, such as social media data, GPS trace data, and historical speed profiles (Shekhar et al., 2014), the traditional methods of visualization do not keep up with the pace and volume of data (Ali, Gupta, Nayak, & Lenka, 2016), innovative visualization analysis tools and techniques are required in the future.
5.3. **DGGS**

The global distribution of spatial vector data will push data management into a global scale in the era of big data. Discrete Global Grid Systems (DGGS), a new OGC standard, is to meet the needs for global sampling, storage, modeling, processing, analysis, and visualization (Purss, Gibb, Samavati, Peterson, & Ben, 2016). The modeling and expressing of big spatial vector data (BSVD) in DGGS is difficult because of its differences and continuous (Tong et al., 2013). However, with globalization and the multi-scale of BSVD, as well as integration and fusion with other data, DGGS will be a good solution.

Apart from above suggestions, in the era of big data, if the security and privacy (Xu et al., 2014), standardization of BSVD are not solved, the sharing and application of BSVD in the future will be hindered largely in the future.

6. **Conclusion**

Eighty percent of the data generated in human life is related to spatial location (Lee & Kang, 2015). Geography has the blessed advantage of big data. Therefore, it is also called the “natural test site” for the research and applications of big data (Wu et al., 2015). Driven by the wave of big data technology, big spatial vector data (BSVD) has been affected and changed, especially for the data management. This paper starts a discussion for the existing work of BSVD management and summarizes three main aspects, namely big spatial data, data storage and organization, data processing, and analysis, which are carried out a detail description from the theoretical and technical levels. The overview of BSVD management is discussed firstly, and then, the big spatial vector data model, storage mode and spatial index are described in the layer of data storage and organization. Furthermore, we discussed the data pre-processing, spatial query, visualization, and spatial analysis. Finally, three future research interests and directions are presented. Big data management is just in its infant stage (Siddiqa et al., 2016), as is big spatial vector data. Therefore, the field needs more theoretical and technical support to better understand and inherit the core of GIS theory and solve the key problems of BSVD effectively. Simultaneously, BSVD should need to comply with the core idea and technology of big data to have a glorious future.

**Data availability statement**

Data sharing is not applicable to this article as no new data were created or analysed in this study.

**Acknowledgments**

The authors would like to thank Zuliang Zhao (PhD student, China Agriculture University) and Sijing Ye (Assistant Professor, Beijing Normal University) for giving helpful suggestions.

**Disclosure statement**

No potential conflict of interest was reported by the authors.
Funding

This work is supported by the Strategic Priority Research Program of Chinese Academy of Sciences [grant number XDA19020201].

ORCID

Xiaochuang Yao http://orcid.org/0000-0001-8068-9415

References

Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., & Saltz, J. (2013). Hadoop-GIS: A high performance spatial data warehousing system over MapReduce. *Proceedings of the VLDB Endowment*, 6(11), 1009–1020.

Al-Badarnah, A., & Al-Alaj, A. (2011). A spatial index structure using dynamic recursive space partitioning. Paper presented at the International Conference on Innovations in Information Technology, Abu Dhabi, United Arab Emirates, April 25–27.

Alarabi, L., & Mokbel, M. F. (2017). A demonstration of ST-Hadoop: A MapReduce framework for big spatio-temporal data. *Proceedings of the VLDB Endowment*, 10(12), 1961–1964. doi:10.14778/3137765.3137819

Alelaiwi, A. (2017). A collaborative resource management for big IoT data processing in Cloud. *Cluster Computing-the Journal of Networks Software Tools and Applications*, 20(2), 1791–1799. doi:10.1007/s10586-017-0839-y

Ali, S. M., Gupta, N., Nayak, G. K., & Lenka, R. K. (2016). Big data visualization: Tools and challenges. Paper presented at the 2nd International Conference on Contemporary Computing and Informatics (IC3I), Noida, India, December 14–17.

Aly, A. M., Mahmoud, A. R., Hassan, M. S., Aref, W. G., Ouzzani, M., Elmeleegy, H., & Qadah, T. (2015). AQWA: Adaptive query workload aware partitioning of big spatial data. *Proceedings of the VLDB Endowment*, 8(13), 2062–2073. doi:10.14778/2831360.2831361

ArcGIS. (2017). GIS tools for Hadoop. ESRI. Retrieved September 25, from https://blogs.esri.com/esri/arcgis/2013/03/25/gis-tools-for-hadoop/

Bellur, U. (2014). On parallelizing large spatial queries using map-reduce. Paper presented at the 13th International Symposium on Web and Wireless Geographical Information Systems, Seoul, South Korea, May 29–30.

Boltcheva, D., & Lévy, B. (2017). Surface reconstruction by computing restricted Voronoi cells in parallel. *Computer-Aided Design*, 90, 123–134. doi:10.1016/j.cad.2017.05.011

Bradley, P. S. (2013). Implications of big data analytics on population health management. *Big Data*, 1(3), 152–159. doi:10.1089/big.2013.0019

Cervone, G., Sava, E., Huang, Q., Schnebele, E., Harrison, J., & Waters, N. (2016). Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 Boulder flood case study. *International Journal of Remote Sensing*, 37(1), 100–124. doi:10.1080/01431161.2015.1117684

Chen, B., Yuan, H., Li, Q., Shaw, S., Lam, W. H. K., & Chen, X. (2015). Spatiotemporal data model for network geographic analysis in the era of big data. *International Journal of Geographical Information Science*, 30(6), 1041–1071. doi:10.1080/13658816.2015.1104317

Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., & Zhou, X. (2013). Big data challenge: A data management perspective. *Frontiers of Computer Science*, 7(2), 157–164. doi:10.1007/s11704-013-3903-7

Chi, M., Plaza, A., Benediktsson, J. A., Sun, Z., Shen, J., & Zhu, Y. (2016). Big data for remote sensing: Challenges and opportunities. *Proceedings of the IEEE*, 104(11), 2207–2219. doi:10.1109/Jproc.2016.2598228

Ding, Z., Chen, Z., & Yang, Q. (2014). IoT-SVKSearch: A real-time multimodal search engine mechanism for the internet of things. *International Journal of Communication Systems*, 27(6), 871–897. doi:10.1002/dac.2647

Eldawy, A., Alarabi, L., & Mokbel, M. F. (2015). Spatial partitioning techniques in SpatialHadoop. *Proceedings of the VLDB Endowment*, 8(12), 1602–1605. doi:10.14778/2824032.2824057
Eldawy, A., Eldanainy, M., Bakeer, A., Abdelmotaleb, A., & Mokbel, M. F. (2015). Sphinx: Distributed execution of interactive SQL queries on big spatial data. Paper presented at the 23rd ACM Sigspatial International Conference on Advances In Geographic Information Systems, Seattle, WA, November 3–6.

Eldawy, A., Li, Y., Mokbel, M. F., & Janardan, R. (2013). CG-Hadoop: Computational geometry in MapReduce. Paper presented at the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Orlando, FL, USA, November 5–8.

Eldawy, A., & Mokbel, M. F. (2013). A demonstration of spatialhadoop: An efficient mapreduce framework for spatial data. Proceedings of the VLDB Endowment, 6(12), 1230–1233. doi:10.14778/2536274.2536283

Eldawy, A., & Mokbel, M. F. (2014). Pigeon: A spatial MapReduce language. Paper presented at the 30th IEEE International Conference on Data Engineering, Chicago, IL, USA, March 31–April 04.

Eldawy, A., & Mokbel, M. F. (2015a). The era of big spatial data. Paper presented at the 31st IEEE International Conference on Data Engineering Workshops, Seoul, South Korea, April 13–17.

Eldawy, A., & Mokbel, M. F. (2015b). SpatialHadoop: A MapReduce framework for spatial data. Paper presented at the 31st IEEE International Conference on Data Engineering, Seoul, South Korea, April 13–17.

Eldawy, A., Mokbel, M. F., Alharthi, S., Alzaidy, A., Tarek, K., & Ghani, S. (2015). SHAHED: A MapReduce-based system for querying and visualizing spatio-temporal satellite data. Paper presented at the 31st IEEE International Conference on Data Engineering, Seoul, South Korea, April 13–17.

Eldawy, A., Mokbel, M. F., & Jonathan, C. (2016). HadoopViz: A MapReduce framework for extensible visualization of big spatial data. Paper presented at the 32nd IEEE International Conference on Data Engineering, Helsinki, Finland, May 16–20.

Fan, J., Ji, M., Gu, G., & Sun, Y. (2014). Optimization approaches to mpi and area merging-based parallel buffer algorithm. Boletim de Ciências Geodésicas, 20(2), 237–256. doi:10.1590/s1982-21702014000200015

Feng, J., Tang, Z., Wei, M., & Xu, L. (2014). HQ-Tree: A distributed spatial index based on Hadoop. China Communications, 11(7), 128–141. doi:10.1109/CC.2014.6895392

Fries, S., Boden, B., Stepient, G., & Seidl, T. (2014). Phidi: Parallel similarity self-join for high-dimensional vector data with mapreduce. Paper presented at the 30th IEEE International Conference on Data Engineering, Chicago, IL, USA, March 31–April 04.

Goodchild, M. F. (2013). The quality of big (geo)data. Dialogues in Human Geography, 3(3), 280–284. doi:10.1177/2043820613513392

Guo, H. (2017). Big Earth data: A new frontier in Earth and information sciences. Big Earth Data, 1(1–2), 4–20. doi:10.1080/20964471.2017.1403062

Guo, H., Wang, L., & Liang, D. (2016). Big Earth Data from space: A new engine for Earth science. Science Bulletin, 61(7), 505–513. doi:10.1007/s11434-016-1041-y

Guo, M., Guan, Q., Xie, Z., Wu, L., Luo, X., & Huang, Y. (2015). A spatially adaptive decomposition approach for parallel vector data visualization of polylines and polygons. International Journal of Geographical Information Science, 29, 1419–1440.

Hameurlain, A., & Morvan, F. (2016). Big Data management in the cloud: Evolution or crossroad?. Paper presented at the 12th International Scientific Conference on Beyond Databases, Architectures and Structures (BDAS), Ustron, Poland, May 31–June 03.

Huang, F., Wen, C., Luo, H., Cheng, M., Wang, C., & Li, J. (2016). Local quality assessment of point clouds for indoor mobile mapping. Neurocomputing, 196, 59–69. doi:10.1016/j.neucom.2016.02.033

Hughes, J. N., Annex, A., Eichelberger, C. N., Fox, A., Hulbert, A., & Ronquest, M. (2015). GeoMesa: A distributed architecture for spatio-temporal fusion. Paper presented at the Geospatial Informatics, Fusion, and Motion Video Analytics V, Baltimore, MD, USA, April 20–21.

Jhummarwala, A., Mazin, A., & Potdar, M. B. (2016). Geospatial Hadoop (GS-Hadoop) an efficient mapreduce based engine for distributed processing of shapefiles. Paper presented at the 2nd International Conference on Advances in Computing, Communication, & Automation, Bareilly, India, September 30–October 1.

Kanchi, S., Sandilya, S., Ramkrishna, S., Manjrekar, S., & Vhadgar, A. (2015). Challenges and solutions in big data management – an overview. Paper presented at the 3rd International Conference on
Future Internet Of Things And Cloud (Ficloud) And International Conference on Open And Big (Obd), Rome, Italy, August 24–26.

Karim, A., Siddiqa, A., Safdar, Z., Razzaq, M., Gillani, S. A., Tahir, H., … Imran, M. (2017). Big data management in participatory sensing: Issues, trends and future directions. Future Generation Computer Systems. doi:10.1016/j.future.2017.10.007

Keim, D., Qu, H., & Ma, K.-L. (2013). Big-Data visualization. IEEE Computer Graphics and Applications, 33(4), 20–21. doi:10.1109/MCG.2013.54

Lee, J.-G., & Kang, M. (2015). Geospatial big data: Challenges and opportunities. Big Data Research, 2(2), 74–81. doi:10.1016/j.bdr.2015.01.003

Li, D. (2016). Towards geo-spatial information science in big data era. Cehui Xuebao/Acta Geodaetica et Cartographica Sinica, 45(4), 379–384. doi:10.11947/j.AGCS.2016.20160057

Li, G., & Huang, Z. (2017). Data infrastructure for remote sensing big data: Integration, management and on-demand service. Jisuanji Yanjiu yu Fazhan/Computer Research and Development, 54(2), 267–283. doi:10.7544/issn1000-1239.2017.20160837

Li, J., Xu, Z., Jiang, Y., & Zhang, R. (2014). The overview of big data storage and management. Paper presented at the 13th IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI-CC), London, UK. August 18–20.

Li, L., Hu, W., Zhu, H., Li, Y., & Zhang, H. (2017). Tiled vector data model for the geographical features of symbolized maps. Plos One, 12(5), 1–26. doi:10.1371/journal.pone.0176387

Li, Q., & Li, D. (2014). Big data GIS. Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics and Information Science of Wuhan University, 39(6), 641–646. doi:10.13203/j.whugis20140150

Li, X., & Zheng, W. (2013). Parallel spatial index algorithm based on hilbert partition. Paper presented at the 5th International Conference on Computational and Information Sciences, Shiyan, Hubei, China. June 21–23.

Lin, W., Zhou, H., & Xia, P. (2016). An effective NoSQL-based vector map tile management approach. Isprs International Journal of Geo-Information, 5(11), 1–25. doi:10.3390/ijgi5110215

Liu, J., Fang, Y., Guo, C., & Gao, K. (2014). Research progress in location big data analysis and processing. Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics and Information Science of Wuhan University, 39(4), 379–385. doi:10.13203/j.whugis20140210

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., … Shi, L. (2015). Social sensing: A new approach to understanding our socioeconomic environments. Annals of the Association of American Geographers, 105(3), 512–530. doi:10.1080/00045608.2015.1018773

Lu, F., & Zhang, H. (2014). Big data and generalized GIS. Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics and Information Science of Wuhan University, 39(6), 645–654. doi:10.13203/j.whugis20140148

Lu, G., Yuan, L., & Yu, Z. (2017). Surveying and mapping geographical information from the perspective of geography. Cehui Xuebao/Acta Geodaetica et Cartographica Sinica, 46(10), 1549–1556. doi:10.11947/j.AGCS.2017.20170338

Lu, J., & Guting, R. H. (2014). Parallel SECONDO: A practical system for large-scale processing of moving objects. Paper presented at the 30th IEEE International Conference on Data Engineering, Chicago, IL, USA. March 31–April 04.

Lu, P, Chen, G., Ooi, B. C., Vo, H. T., & Wu, S. (2014). ScalaGiST: Scalable generalized search trees for mapreduce systems [innovative systems paper]. Proceedings of the VLDB Endowment, 7(14), 1797–1808. doi:10.14778/2733085.2733087

Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., & Jie, W. (2015). Remote sensing big data computing: Challenges and opportunities. Future Generation Computer Systems, 51, 47–60. doi:10.1016/j.future.2014.10.029

Magdy, A., Mokbel, M. F., Elnikety, S., Nath, S., & He, Y. (2016). Venus: Scalable real-time spatial queries on microblogs with adaptive load shedding. IEEE Transactions on Knowledge & Data Engineering, 28(2), 356–370. doi:10.1109/TKDE.2015.2493531

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. Analytics.

Mapbox. (2018). Awesome implementations of the vector tile. Author. Retrieved January 5, from https://github.com/mapbox/awesome-vector-tiles
Mcafee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, 90(10), 60–66.

McCoy, M. D. (2017). Geospatial big data and archaeology: Prospects and problems too great to ignore. *Journal of Archaeological Science*, 84, 74–94. doi:10.1016/j.jas.2017.06.003

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *Geojournal*, 80(4), 449–461. doi:10.1007/s10708-014-9602-6

Mulyono, S., & Fanany, M. I. (2015). Remote sensing big data utilization for paddy growth stages detection. Paper presented at the IEEE International Conference on Aerospace Electronics and Remote Sensing Technology, Bali, Indonesia, December 3–5.

Nishimura, S., Das, S., Agrawal, D., & El Abbadi, A. (2011). MD-HBase: A scalable multi-dimensional data infrastructure for location aware services. Paper presented at the 12th IEEE International Conference on Mobile Data Management, Lulea, Sweden, June 6–9.

Nobari, S., Tauheed, F., Heinis, T., Karras, P., Bressan, S., & Ailamaki, A. (2013). TOUCH: In-memory spatial join by hierarchical data-oriented partitioning. Paper presented at the Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, June 22–27.

Pekturk, M. K., & Unal, M. (2017). A review on real-time big data analysis in remote sensing applications. Paper presented at the 25th Signal Processing and Communications Applications Conference (SIU), Antalya, Turkey, May 15–18.

Puri, S., & Prasad, S. K. (2014). Output-sensitive parallel algorithm for polygon clipping. Paper presented at the 43rd International Conference on Parallel Processing (ICPP), Minneapolis, MN, USA, September 9–12.

Puri, S., & Prasad, S. K. (2015). A parallel algorithm for clipping polygons with improved bounds and a distributed overlay processing system using MPI. Paper presented at the 15th IEEE ACM International Symposium on Cluster Cloud and Grid Computing, Shenzhen, China, May 4–7.

Purss, M. B. J., Gibb, R., Samavati, F., Peterson, P., & Ben, J. (2016). The Ogc (R) discrete global grid system core standard: A framework for rapid geospatial integration. Paper presented at the 36th IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, July 10–15.

Ray, S., Simion, B., Brown, A. D., & Johnson, R. (2013). A parallel spatial data analysis infrastructure for the cloud. Paper presented at the Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Orlando, FL, USA, November 5–8.

Shekhar, S., Evans, M. R., Gunturi, V., Yang, Y., & Cugler, D. C. (2014). Benchmarking spatial big data. Paper presented at the 2nd Workshop on Specifying Big Data Benchmarks, Pune, India, December 17–18.

Shen, D., Wong, D. W., Camelli, F., & Liu, Y. (2013). An ArcScene plug-in for volumetric data conversion, modeling and spatial analysis. *Computers & Geosciences*, 61, 104–115. doi:10.1016/j.cageo.2013.08.004

Siddiqi, A., Hashem, I. A. T., Yaqoob, I., Marjani, M., Shamshirband, S., Gani, A., & Nasaruddin, F. (2016). A survey of big data management: Taxonomy and state-of-the-art. *Journal of Network and Computer Applications*, 71, 151–166. doi:10.1016/j.jnca.2016.04.008

Stantic, B., & Pokorný, J. (2014). Opportunities in big data management and processing. *Frontiers in Artificial Intelligence & Applications*, 270, 15–26. doi:10.3233/978-1-61499-458-9-15

Starinshak, D. P., Owen, J. M., & Johnson, J. N. (2014). A new parallel algorithm for constructing Voronoi tessellations from distributed input data. *Computer Physics Communications*, 185(12), 3204–3214. doi:10.1016/j.cpc.2014.08.020

Storey, V. C., & Song, I.-Y. (2017). Big data technologies and management: What conceptual modeling can do. *Data & Knowledge Engineering*, 108, 50–67. doi:10.1016/j.datak.2017.01.001

Sui, D., Ye, X., & Gan, T. (2014). Open GIS for big data: Opportunities and impediments. *Progress in Geography*, 33(6), 723–737. doi:10.11820/dlkxjz.2014.06.001
Sun, Z., Chen, F., Chi, M., & Zhu, Y. (2015). A spark-based big data platform for massive remote sensing data processing. Paper presented at the 2nd International Conference on Data Science (ICDS), Univ Technol Sydney, Sydney, Australia, August 8–9.

Taleb, I., Dssouli, R., & Serhani, M. A. (2015). Big data pre-processing: A quality framework. Paper presented at the IEEE International Congress on Big Data, New York, NY, USA, June 27–July 2.

Tan, H., Luo, W., & Ni, L. M. (2012). CloST: A hadoop-based storage system for big spatio-temporal data analytics. Paper presented at the Proceedings of the 21st ACM international conference on Information and knowledge management, Maui, Hawaii, USA, October 29–November 2.

Tang, W., & Feng, W. (2017). Parallel map projection of vector-based big spatial data: Coupling cloud computing with graphics processing units. *Computers, Environment and Urban Systems, 61*, 187–197. doi:10.1016/j.compenvurbsys.2014.01.001

Tong, X., Ben, J., Liu, Y., & Zhang, Y. (2013). Modeling and expression of vector data in the hexagonal discrete global grid system. *Isprs Webmaps 2013 & Dmgis 2013 Topics: Global Spatial Grid & Cloud-Based Services 40-4-W2, 15–25*. doi:10.5194/isprsarchives-XL-4-W2-15-2013

Tsou, M.-H. (2015). Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science, 42*(Suppl 1), 70–74. doi:10.1080/15230406.2015.1059251

Wang, H., Guan, X., & Wu, H. (2017). A hybrid parallel spatial interpolation algorithm for massive LiDAR point clouds on heterogeneous CPU-GPU systems. *ISPRS International Journal of Geo-Information, 6*(11), 363. doi:10.3390/ijgi6110363

Wang, J. (2017). Cartography in the age of spatio-temporal big data. *Cehui Xuebao/Acta Geodaetica et Cartographica Sinica, 46*(10), 1226–1237. doi:10.11947/j.AGCS.2017.20170308

Wang, L., Chen, B., & Liu, Y. (2013). Distributed storage and index of vector spatial data based on HBase. Paper presented at the 21st International Conference on Geoinformatics, Kaifeng, China, June 20–22.

Wang, L., Geng, H., Liu, P., Lu, K., Kolodziej, J., Ranjan, R., & Zomaya, A. Y. (2015). Particle Swarm optimization based dictionary learning for remote sensing big data. *Knowledge-Based Systems, 79*, 43–50. doi:10.1016/j.knosys.2014.10.004

Wang, L., Ma, Y., Yan, J., Chang, V., & Zomaya, A. Y. (2018). pipsCloud: High performance cloud computing for remote sensing big data management and processing. *Future Generation Computer Systems, 78*, 353–368. doi:10.1016/j.future.2016.06.009

Wang, Y., Liu, Z., Liao, H., & Li, C. (2015). Improving the performance of GIS polygon overlay computation with MapReduce for spatial big data processing. *Cluster Computing-the Journal of Networks Software Tools and Applications, 18*(2), 507–516. doi:10.1007/s10586-015-0428-x

Wei, H., Du, Y., Liang, F., Zhou, C., Liu, Z., Yi, J., … Wu, D. (2015). A k-d tree-based algorithm to parallelize Kriging interpolation of big spatial data. *GIScience & Remote Sensing, 52*(1), 40–57. doi:10.1080/15481603.2014.1002379

Whitman, R. T., Park, M. B., Ambrose, S. M., & Hoel, E. G. (2014). Spatial indexing and analytics on hadoop. Paper presented at the Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, New York, NY, USA, November 4–7.

Wojda, P., & Brouyère, S. (2013). An object-oriented hydrogeological data model for groundwater projects. *Environmental Modelling & Software, 43*, 109–123. doi:10.1016/j.envsoft.2013.01.015

Wu, Z., Chai, Y., Dang, A., Gong, J., Gao, S., Yue, Y., … Liu, Y. (2015). Geography interact with big data: Dialogue and reflection. *Geographical Research, 12*(34), 2207–2221. doi:10.11821/dlyj201512001

Xie, D., Li, F., Yao, B., Li, G., Chen, Z., Zhou, L., & Guo, M. (2016). Simba: Spatial in-memory big data analysis. Paper presented at the 24th ACM SIGSPATIAL International Conference on Advances In Geographic Information Systems, San Francisco, CA, USA, October 31–November 3.

Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: Privacy and data mining. *IEEE Access, 2*, 1149–1176. doi:10.1109/access.2014.2362522

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2016). Big DATA and cloud computing: Innovation opportunities and challenges. *International Journal of Digital Earth, 10*(1), 13–53. doi:10.1080/17538947.2016.1239771

Yang, C., Raskin, R., Goodchild, M., & Gahegan, M. (2010). Geospatial cyberinfrastructure: Past, present and future. *Computers, Environment and Urban Systems, 34*(4), 264–277. doi:10.1016/j.compenvurbsys.2010.04.001
Yang, C., Yan, X., & Nebert, D. (2013). Redefining the possibility of digital Earth and geosciences with spatial cloud computing. *International Journal of Digital Earth, 6*(4), 297–312. doi:10.1080/17538947.2013.769783

Yang, C., Yu, M., Hu, F., Jiang, Y., & Li, Y. (2017). Utilizing cloud computing to address big geospatial data challenges. *Computers, Environment and Urban Systems, 61*, 120–128. doi:10.1016/j.compenvurbsys.2016.10.010

Yao, X., Mokbel, M. F., Alarabi, L., Eldawy, A., Yang, J., Yun, W., … Zhu, D. (2017). Spatial coding-based approach for partitioning big spatial data in Hadoop. *Computers & Geosciences, 106*, 60–67. doi:10.1016/j.cageo.2017.05.014

Yao, X., Yang, J., Li, L., Yun, W., Zhao, Z., Ye, S., & Zhu, D. (2017). LandQv1: A GIS cluster-based management information system for arable land quality big data. Paper presented at the 6th International Conference on Agro-Geoinformatics, Fairfax, VA, USA, August 7–10.

Yao, X., Zhu, D., Ye, S., Zhang, N., & Li, L. (2014). Spatial interpolation methods study based on geostatistics for the grasshopper population. *Sensor Letters, 12*, 645–650. doi:10.1166/sl.2014.3095

Yao, X., Zhu, D., Yun, W., Peng, F., & Li, L. (2017). A WebGIS-based decision support system for locust prevention and control in China. *Computers and Electronics in Agriculture, 140*, 148–158. doi:10.1016/j.compag.2017.06.001

Ye, S., Yan, T., Yue, Y., Lin, W., Li, L., Yao, X., … Zhu, D. (2016). Developing a reversible rapid coordinate transformation model for the cylindrical projection. *Computers & Geosciences, 89*, 44–56. doi:10.1016/j.cageo.2016.01.007

You, S., Zhang, J., & Gruenwald, L. (2015a). Large-scale spatial join query processing in cloud. Paper presented at the 31st IEEE International Conference on Data Engineering Workshops, Seoul, South Korea, April 13–17.

You, S., Zhang, J., & Gruenwald, L. (2015b). Spatial join query processing in cloud: Analyzing design choices and performance comparisons. Paper presented at the 44th International Conference on Parallel Processing Workshops, Beijing, China, September 1–4.

Yu, E. G., Di, L., Rahman, M. S., Lin, L., Zhang, C., Hu, L., … Yang, G. (2017). Performance improvement on a Web Geospatial service for the remote sensing flood-induced crop loss assessment web application using vector tiling. Paper presented at the 6th International Conference on Agro-Geoinformatics, Fairfax, VA, USA, August 7–10.

Yu, J., Wu, J., & Sarwat, M. (2015). GeoSpark: A cluster computing framework for processing large-scale spatial data. Paper presented at the Proceedings of the 23rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, Washington, November 3–6.

Zhang, G. (2017). Construction of the college education quality management system based on big data and its evaluation. *Agro Food Industry Hi-Tech, 28*(1), 3124–3127.

Zhang, J., & You, S. (2012). Speeding up large-scale point-in-polygon test based spatial join on GPUs. Paper presented at the Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, Redondo Beach, CA, USA, November 6.

Zhang, X. (2014). Spatial analysis in the era of big data. *Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics and Information Science of Wuhan University, 39*(6), 655–659. doi:10.13203/j.whugis20140143

Zhao, L., Chen, L., Ranjan, R., Choo, K.-K. R., & He, J. (2015). Geographical information system parallelization for spatial big data processing: A review. *Cluster Computing, 19*(1), 139–152. doi:10.1007/s10586-015-0512-2

Zheng, K., & Fu, Y. (2013). Research on vector spatial data storage schema based on Hadoop platform. *International Journal of Database Theory and Application, 6*(5), 85–94. doi:10.14257/ijdta.2013.6.5.08

Zheng, Y. (2015). Introduction to urban computing. *Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics and Information Science of Wuhan University, 40*(1), 1–13. doi:10.13203/j.whugis20140718

Zheng, Y. (2017). Urban computing: Enabling urban intelligence with big data. *Frontiers of Computer Science, 11*(1), 1–3. doi:10.1007/s11704-016-6907-2

Zhong, R., Li, G., Tan, K.-L., Zhou, L., & Gong, Z. (2015). G-Tree: An efficient and scalable index for spatial search on road networks. *IEEE Transactions on Knowledge and Data Engineering, 27*(8), 2175–2189. doi:10.1109/tkde.2015.2399306
Zhong, Y., Fang, J., & Zhao, X. (2013). VegaIndexer: A distributed composite index scheme for big spatio-temporal sensor data on cloud. Paper presented at the 33rd IEEE International Geoscience and Remote Sensing Symposium, IGARSS, Melbourne, VIC, Australia, July 21–26.

Zhong, Y., Han, J., Zhang, T., Li, Z., Fang, J., & Chen, G. (2012). Towards parallel spatial query processing for big spatial data. Paper presented at the 26th IEEE International Parallel and Distributed Processing Symposium Workshops, Shanghai, China, May 21–25.

Zhou, Y., Zhou, C., Ma, T., Han, J, Xu, T., & Ji, M. (2015). A double-index and data divide-conquer based parallel point-polygon overlay method. Geography and Geo-Information Science, 31(02), 1–6. doi:10.3969/j.issn.1672-0504.2015.02.001

Zhu, X., Huo, J., & Qiu, Q. (2015). A novel methodology for parallel spatial overlay over vector data-A case study with shape file. Paper presented at the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, July 26–31.

Zhuang, Y., Fong, S., Yuan, M., Sung, Y., Cho, K., & Wong, R. K. (2017). Location-based big data analytics for guessing the next Foursquare check-ins. The Journal of Supercomputing, 73(7), 3112–3127. doi:10.1007/s11227-016-1925-2