IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 24, 2020, accepted December 28, 2020, date of publication January 12, 2021, date of current version January 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3051084

Review on Integrating Geospatial Big Datasets and Open Research Issues

SOHAIB AL-YADUMI[1], TAN EE XION[2], SHARON GOH WEI WEI[1], AND PATRICE BOURSIER[2]

[1]School of Computer Science and Engineering, Taylor's University, Subang Jaya 47500, Malaysia
[2]Life Sciences, School of Pharmacy, International Medical University, Bukit Jalil 57000, Malaysia

Corresponding author: Sohaib Al-Yadumi (sohaibmohamedabdullahalyadumi@sd.taylors.edu.my)

This work was supported in part by the Taylor's University under Grant TUFR/2017/004/03.

ABSTRACT Big data and geographic information systems (GIS) are two technologies that have increasingly influenced many areas in the last 10 years and will continue to improve and help solve serious global problems, such as consequences of climate change or global pandemics. A wide spectrum of GIS applications interacts with the continuous growth of geospatial big data sources to drive precise and informed decisions. Geospatial big data integration is designed to accomplish the compatibility of distinct geospatial datasets regardless of their spatial coverage. The large number of geospatial big data sources demand effective data integration for storing and handling such datasets, which will be used for geospatial data analysis and visualization. For instance, risk management datasets related to healthcare and the environment are heterogeneous and disparate. Obtaining a unified view of such geospatial big datasets is complicated and challenging, especially if we consider problems related to healthcare pandemics and environmental disasters. Hence, before we can attempt to predict and mitigate processes occurring in these domains, we must realize that geospatial big data integration is crucial in consolidating datasets. We explore and discuss issues involved in integrating geospatial big datasets in this study. We then classify big data integration processes into three categories, namely, data warehousing, data transformation and integration methods. Furthermore, several research challenges focused on geospatial big data, big earth data, data warehousing, data transformation and linked data are presented. Lastly, open research issues and emerging trends that require in-depth investigations in the near future are highlighted in this study.

INDEX TERMS Big data integration, geographic information system (GIS), geospatial big data.

I. INTRODUCTION

Geographic information systems (GISs) interact with a large number of geospatial big data sources with different formats. GIS can import, export, store, manage, analyze, process and visualize spatial georeferenced data and plays a key role in integrating and analyzing large amounts of geospatial data [1]. A considerable portion of information is referred to as geospatial data, which are collected using technologies, such as global position systems (GPS), radio-frequency identifications, volunteered geographic information and location-based social networks. Geospatial data are used in applications related to land use, environmental management, healthcare, tourism, marketing and many others. However, the majority of these data are only available in isolated diverse data sources and have many data quality issues due to heterogeneous data types and contributors [2].

Pooling geospatial data from different data sources is commonly applied and important in healthcare- and environment-related applications to obtain new knowledge and make informed decisions [3], [4]. An increasing number of geospatial data sources is stored in databases that consist of sensor data, volunteered geographic information and location-based data. Even regular people can create new or update existing geospatial data in near-real time with advanced web technologies and location-based devices and services. Integrating proper geospatial and non-geospatial data can help provide correct and timely information for relevant people because the heterogeneous information will be pooled from multiple sources.

Synthesis of data located in distinct sources is known as data integration [5]. The underlying principle of an

The associate editor coordinating the review of this manuscript and approving it for publication was Genoveffa Tortora.

integration system is the homogeneous perspective of one access interface for data stored and accessed from various data sources through a universal plan in a mediator or data warehouse (DW) [6]. Data integration is generally classified into six categories, namely, manual integration, provision of a shared user interface to users, creation of an integration application, use of a middleware interface, development of homogeneous data access, and creation of shared data storage [7].

Modern decision-making systems rely heavily on DWs, a concept first introduced in the 1980s [8]. Combined data from different sources are stored and made accessible in multidimensional formats in DWs for investigations intended to assist users in improving their knowledge about their business. An extract-transform-load (ETL) procedure is typically applied to collect the majority of DW information from corporate operational databases and involves the extraction and conversion of data into a multidimensional format before loading into the DW as cubes for subsequent analysis via reporting and online analytical processing (OLAP) tools. Rapid bulk-loading methods are generally used to conduct ETL regularly during a time interval of DW inactivity [9].

Although collecting general data from various database systems is difficult, numerous solutions have been proposed for data integration from distinct relational systems. However, the lack of homogeneity in source database systems indicates that dissimilar data models, such as relation or various nonrelational data models, are used [7]. Conventional approaches to data access, discovery and integration are revolutionized by linked data [10]. Geospatial data can be efficiently shared and discovered in spatial data infrastructures (SDI) based on properties of linked data, such as the common data model, standardized mechanism of data access and data discovery based on links. Semantic interoperability amongst various web applications and services can be achieved by addressing the issue of semantic heterogeneity on the basis of ontologies [6].

The issues associated with getting data from a vast number of sources are a challenge. Many differences between conventional and big data integration (BDI) are related to the number of data sources, structural homogeneity, changing nature and highly varied data sources in terms of their quality, such as timeliness, precision and coverage [11]. Additionally, geospatial big data integration is designed to accomplish the compatibility of distinct geospatial datasets regardless of their spatial coverage [12]. Data undergo conversion from a range of formats, projections or systems of reference, followed by adaptation according to a given data model [13]. This process makes it easier for the integrated geospatial data to be analyzed, processed and visualized. This is achieved from the merging of data gathered from various sources with the different methods into a unified view.

Notably, numerous applications ranging from transport planning to crisis risk management depend on BDI for data access and analysis [14]. Many studies in the literature reviewed geospatial big data. For instance, Eldawy and Mokbel [15] presented the era of geospatial big data. Li *et al.* [2] reviewed geospatial big data methods and major challenges. Yao and Li [16] introduced big geospatial vector data management. However, these studies ignored the holistic integration of geospatial big data. Hence, this article reviews the literature on general data and big geospatial integration and aims to develop a common taxonomy that can help researchers understand this field and develop holistic geospatial BDI methods. Furthermore, we believe that this review will be valuable to novice researchers from different fields and domains. An extensive review of existing geospatial data integration methods and a comprehensive investigation of BDI in geospatial environments in the past five years are presented in this study. The extensive review explores existing definitions and characteristics of geospatial and big data integration. The relationship amongst big data, BDI and geospatial data is also discussed. Existing BDI studies are classified into the following categories: (i) data warehousing, (ii) data transformation, and (iii) integration methods. Furthermore, research challenges, several open research issues, and trends on geospatial data are discussed.

The whole paper is organized as follows: Definitions and characteristics of geospatial big data are introduced in Section II. The classification of BDI is presented in Section III. A summary of current research challenges and open research issues, especially problems related to geospatial BDI, is provided in Sections IV and V, respectively. Trends in future development are presented in Section VI. Lastly, the conclusion of the study with discussion and future work is presented in Section VII.

II. DEFINITION AND CHARACTERISTICS OF GEOSPATIAL BIG DATA

According to Liu *et al.* [17], common characterizations of big data are volume, variety and velocity, which are collectively known as the 3V model. IBM [18] considered veracity an important dimension in accurately characterizing big data. Volume refers to different data sources, including sensor and social networks, which generate large amounts of daily data beyond the processing capability of traditional databases. Variety is concerned with the different formats of generated data, including structured and unstructured data. Examples include legacy systems, blog posts, tweets and mobile applications. Tria *et al.* [19] noted that velocity came from the need for rapidly transferring data between sources to remain competitive.

Geospatial data includes position (e.g. building geometry or coordinates) and lexical (e.g. building names) information [2]. A range of modalities is used to capture geospatial data, including GPS, satellite images, social networking, location-based services and high-resolution remote sensors, which are then entered into a GIS database, which stores the georeference-related data that specify spatial information related to connections amongst data points, non-geospatial (attribute) characteristics and other issues. Spatially-located datasets are valuable because they determine not only

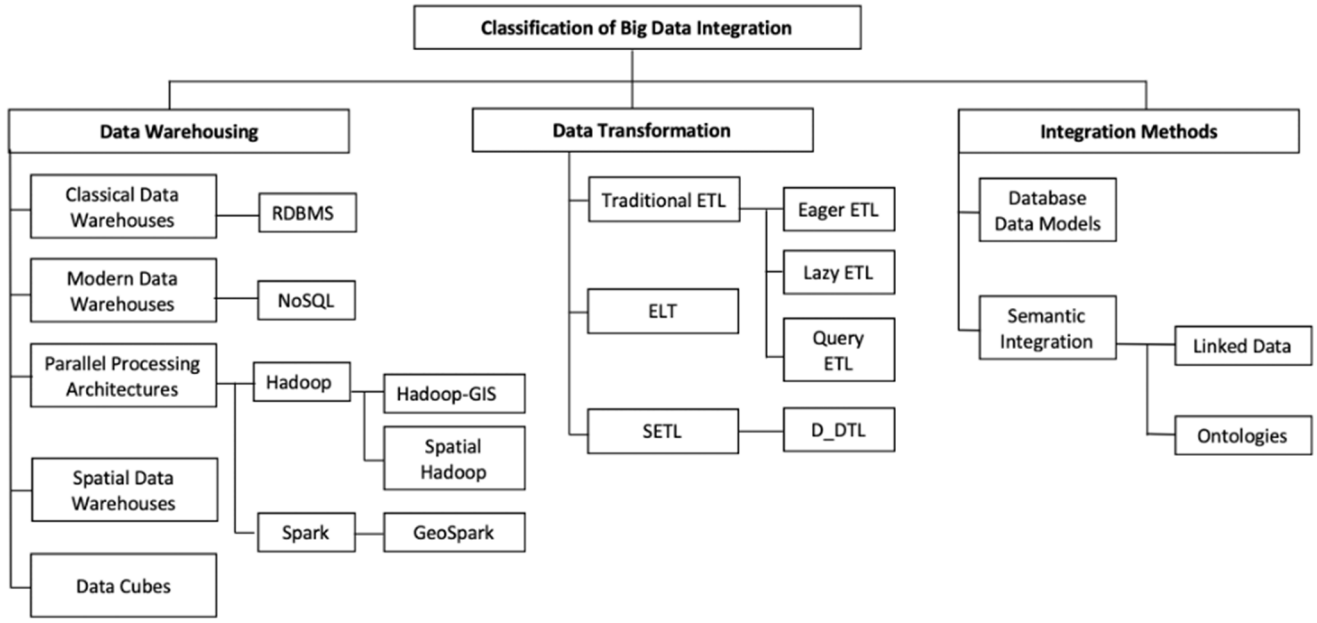

FIGURE 1. Classification of big data integration.

locational features of any given data point but also indicate the timing and type of event. Furthermore, this data can be represented in either raster or vector data type. The latter uses geometrical shapes, such as points and lines in the vector model to represent geospatial elements and has the advantages of accuracy, low volume and high quality when considered with respect to the raster data model [1], [16].

Geospatial data has long been considered as big data. Managing geospatial data is highly important since a great proportion of data is available for geo-referencing, as reflected by the questionable claim that ''80% of data is geographic'' [2]. Nowadays, geospatial big data are valuable in many fields such as data analytics and discovery. Numerous societal applications, such as environmental changes and disaster risk management, can possibly benefit from geospatial big data [20].

III. CLASSIFICATION OF BIG DATA INTEGRATION

Apart from structured information, semi-structured and unstructured data have also been increasingly used by organizations to improve their business analytics-related decision making [21]. Consequently, data from a wide range of sources must be accessed, analyzed and shared. Interoperating and integrating tasks of various information systems are challenging due to the volume, variety and velocity properties of big data and the lack of intersystem uniformity emphasis in big data. The demand for data integration by organizations and enterprises has led to the emergence of a new research field called BDI [11], which is limited by key features of big data.

Hence, BDI plays a key role in consolidating various datasets. In this review, we introduce BDI classification. This classification will help geospatial data players to obtain a holistic comprehension on how to combine disparate geospatial or non-geospatial sources into meaningful and high-value information. Fig. 1 illustrates the main categories of BDI, namely, (i) data warehousing, (ii) data transformation, and (iii) integration methods. Fig. 2 presents the keyword analysis of reviewed articles using VOSviewer software [22]. Big data, data integration, semantic web and DW are commonly repeated keywords.

A. DATA WAREHOUSING

DW is created within environments that were previously dependent on traditional servers used to run relational databases. Although the term was used in the 1980s, data warehousing became a standalone research topic in the late 1990s. Notably, DW has stringent connections with related topics, such as data visualization and database integration [23]. DW was mainly invented due to the need for storing and querying historical data and is a means of extracting scattered important information saved in various information systems and collating it in a centralized and integrated storage system. A notable shift towards the use of parallel architectures to cater to the demanding aspects of big data requirements has been observed. Santoso and Yulia [24] noted that various architecture approaches, such as shared-disk, shared-memory, shared-nothing and shared-everything methods, can be adopted. DW is generally classified into five types, namely, classical, modern, parallel processing architecture, spatial DWs, and data cubes. Table 1 lists recent studies related to DW and big data according to their publication year. The last column (big data category) shows the typically used big data characterization in the reviewed studies.

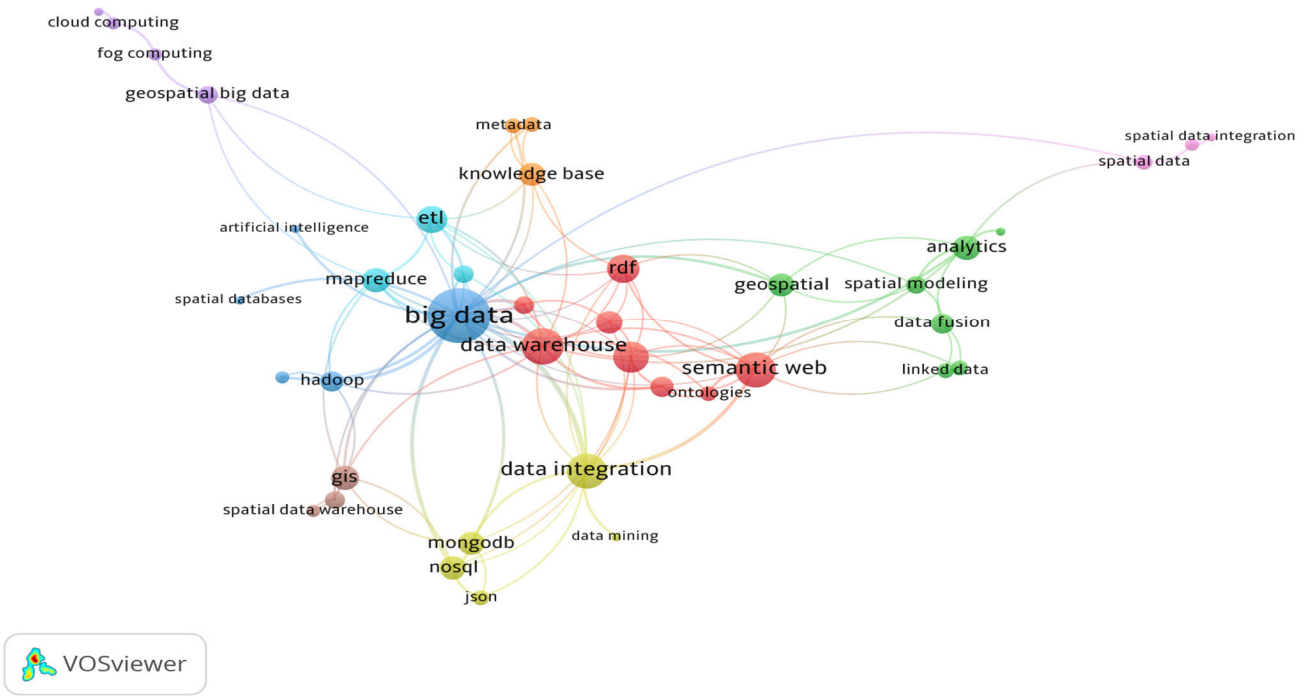

FIGURE 2. Frequent keywords in geospatial big data articles from 2015 to 2020.

Amongst the big data categories, variety and volume are commonly utilized in nearly all the studies. Velocity and veracity are neglected in most studies because of the lack of holistic solutions due to the nature of big data.

1) CLASSICAL DATA WAREHOUSES

Enterprises commonly use different IT systems to cover different functions and business areas, such as customer relationship management, accounting and human resources. DWs are generally created when management requires an integrated and streamlined strategy for accessing information [25]. A classical DW is built from a logical collection of data that provides a summary of the organization's information but stands separate from any other operational databases to allow the integration of many different datasets regardless of the application or system the information is obtained from. Centralization creates a one-door access policy for management to obtain and assess information for strategic decision making. However, the traditional DW is incapable of handling large volumes of information with the arrival of big data. Santoso and Yulia [24] noted that data processing must be conducted in parallel. Describing data are structured in the classical DW, which is incapable of dealing with semi-structured or unstructured data. Hence, modern DWs can be an option for large volumes of datasets.

2) MODERN DATA WAREHOUSES

Modern DWs are designed to deal with emerging data sources, including sensor and social networks, which generate large amounts of data on a daily basis. Not only structured query language (NoSQL) is equipped to process high volumes of structured, semi-structured and unstructured data. Bicevska and Oditis [25] demonstrated that NoSQL technology is suitable for processing high volumes of data from various settings. Akid and Ayed [26] extended the NoSQL technology and proposed an entirely new architecture of modern DW to construct a graph document-oriented NoSQL database for storing and analyzing big social data. However, both [25] and [26] ignored the ETL process requirement under the NoSQL DW, thereby limiting the creation of these DWs and hindering the transfer of required data.

3) PARALLEL PROCESSING ARCHITECTURES

Massive data are generated by a number of devices, such as Internet of Things (IoT) and location-based social networks. Parallel processing architectures (e.g. MapReduce) [27] can solve classical DW issues, which cannot be addressed by single-node solutions because of the high load required for data analysis. Analysis of big data in distributed environments is typically conducted via the MapReduce-based Hadoop frameworks [28]. Supporting effective queries on large amounts of geospatial data is increasingly important in a wide

TABLE 1. Summary of data warehouse studies related to big data and limitations.

Article	Publication Year	Description	Limitations/Drawbacks	Data Type/ Database Model	Big Data Category
Moulai [43]	2018	The information warehousing model was introduced in this study as an alternative to the current DW paradigm by separating the three key entities of knowledge, information and data. A common and original information warehouse architecture was proposed for both the storage and analysis of information from a wide range of sources, including social media, press articles and scientific journals.	- Only limited to and valid for specific domains, such as social media, press articles and scientific papers - Unable to support geospatial data and parallel processing	RDBMS NoSQL	Variety
Mazzei and Guida [44]	2018	A SOLAP and SDW prototype was created in this work. The finished project illustrated the possibility of providing users with an analysis tool containing high volumes of information from a wide variety of data sources.	- Unable to support distributed architectures and parallel processing	RDBMS GIS formats	Variety
Goyal et al. [1]	2017	GIS data models with geospatial data mining strategies were integrated in this study. A big data approach for integrating GIS datasets was proposed to achieve informative data analysis results.	- Analyzing and mining of forestry data are difficult to apply - Forestry data mining process takes a long time to complete - Requires additional computing powers	MapReduce Geospatial data	Volume Variety
Akid and Ayed [26]	2017	An entirely new architecture was proposed in this study to construct a graph DW for storing and analysing big social data. The system uses a document-oriented NoSQL database management system to cater specifically to the storage needs of data shared on social platforms, and nodes and links required for analysis are stored in a graph NoSQL database.	- Applicable to and valid only for social networking sources - Lacks a standard query language	NoSQL	Volume
Bimonte et al. [45]	2017	A conceptual model that integrates regular grids of points into SOLAP was provided in this study by extending logical models of SDW in many ways. The proposed FieldMDX, an extension of the existing MDX model, can generate the continuity of an incomplete field using geospatial interpolation functions.	- Unable to support distributed architectures, such as NoSQL - Excludes cloud-based solutions	RDBMS Geospatial data	Volume Variety
Barkhord and Niamanesh [46]	2017	Atrak is a technique proposed for building a MapReduce DW. This approach solves the large dimension problem caused by a massive amount of data and allows every node to work individually. Hence, each node has the ability to run its queries separately.	- Big data volume will affect the latency and processing time - Some nodes still require linking parts of data	RDBMS MapReduce	Volume Variety
Bicevska and Oditis [25]	2017	Potential outcomes were presented in this study by applying the NoSQL database as the new DW. Implementation aspects and challenges needed for this approach were outlined.	- The ETL process is missing in implementation in the NoSQL DW	NoSQL	Volume Variety
Eldawy and Mokbel [47]	2015	SpatialHadoop was developed to represent a complete MapReduce model that intrinsically supports geospatial data. The Hadoop code base encompasses four layers of SpatialHadoop (i.e. language, storage, MapReduce and operations).	- Unable to support real-time processing - Applies static data partitioning technique between slave machines - Requires optimization of geospatial queries	MapReduce Geospatial data	Volume Variety

range of application domains. Rapidly answering queries is a key condition for geospatial applications involving large amounts of data. This condition necessitates a scalable configuration capable of large-scale querying of geospatial data. For this purpose, the spatial data warehousing (SDW) system Hadoop-GIS that demonstrates scalability and efficiency was proposed in [29] as an option to run large-scale geospatial queries on Hadoop. Furthermore, GeoSpark [30], an in-memory cluster computing system, was developed to process large-scale geospatial data and support various forms of geospatial data, indices and operations by expanding the core of Apache Spark; hence, GeoSpark expands the concept of resilient distributed datasets. Compared with Hadoop-based systems, GeoSpark has been empirically demonstrated to improve run-time performance [15].

4) SPATIAL DATA WAREHOUSES

SDW is utilized by geospatial decision support systems (DSSs) to query and analyze position information-related data [13], [31]. According to Baazaoui-Zghal [32], SDW has become increasingly necessary in many fields, such as

healthcare and environmental disaster management. SDW is generally considered an efficient option when large volumes of data are involved. Decision makers have become increasingly aware of the urgent need for effectively storing continuously growing amounts of both structured and unstructured data. Li *et al.* [33] suggested the use of a NoSQL database as a warehouse for geospatial big data whilst using a traditional geospatial database as the application server. Baazaoui-Zghal [32] proposed an ontology-based fuzzy SDW for contextual search and recommendation based on the integration of uncertain data at multiple levels of the knowledge layer whilst decisional architecture, contextual search and recommendation remain the same. The extraction of relevant and interesting information from SDWs can be complicated. Hence, recommendation systems aim to aid users in their navigation through large datasets and finding of relevant information based on their analytical objectives and personal preferences.

5) DATA CUBES

Constituting extensive multi-dimensional arrays [34], data cubes are among the newest and most effective approaches for Big Earth Data (BED) storage and analysis. Based on these data cubes, information from gridded data kept spatial, temporal, and various other dimensions can be trimmed, sliced or extracted more rapidly [35]. Besides affording a number of advances, data cubes also promote the application of the standards of the open geospatial consortium (OGC) in interactions with geospatial data. Several important endeavors have been initiated in recent times to resolve the big data problems facing various scientific groups based on the use of data cubes. For example, EarthServer4 ensures compatibility between BED analysis and a range of integrated products [36], while earth observation (EO) data can be better organized and analyzed on the basis of the data structures and instruments encompassed in the analytical framework of Open Data Cube [37]. Furthermore, the continuous gathering and analysis of cutting-edge data cubes and array databases (e.g. standards and implementations) are facilitated by the Research Data Alliance [35]. Nevertheless, the use of data cubes is problematic because it demands novel storage and processing paradigms to ensure the same speed of queries along each dimension [34].

B. DATA TRANSFORMATION

The ETL process can be used to integrate and load data from different sources completely; this process begins with the creation of an integrated repository of data [14]. Tasks involved in the ETL process are data acquisition from more than one source (extraction), data processing according to the warehouse integrity standards (transformation), and data populating in the warehouse in the form of new records (loading) [38]. Thus, the ETL process is crucial in the creation of DW [39] because it loads all the existing data in the DW prior to the initiation of user queries [9]. The ETL process is classified into three types, namely, traditional ETL, extract-load-transform (ELT) and spatial ETL (SETL). Table 2 presents the summary of some studies related to ETL tools and our highlighted advantages and disadvantages. Many studies lack the support for extracting and transforming data, and use NoSQL databases, which are unable to handle unstructured geospatial data.

1) TRADITIONAL ETL

Traditional ETL is called an eager process because it starts by loading complete data sources in an integrated repository of data [14]. Data are ineffectively managed through the eager ETL mainly because of (1) the high loading time and bandwidth, and (2) data authorization which could be circumvented if all the data source contents are loaded in an integrated data repository; however, this bypass will increase the risk of data breach by allowing data access to users [14]. A number of approaches has been proposed to solve these issues [9], [40]. Lazy ETL involves the eager integration and the loading solely of metadata rather than the actual data entry [14]. Traditional ETL techniques are graphical user interface (GUI) to ease the process of moving data from the source to the target system and hand-coded ETL system, which provides high adaptability. Many open-source and commercial ETL tools, such as Talend, Informatica and IBM, are available. Iswas *et al.* [34] proposed an ETL approach that supports near real-time ETL processes by applying incremental loading. However, this model is unable to support GUI and requires users to work with a few lines of code. Sreemathy *et al.* [42] used Talend Open Studio as a GUI ETL tool to ease the data integration and analysis process. GUI ETL tools are very helpful because they provide users with many ready-to-use features and can be implemented by different fields due to the repetitive process of ETL.

2) ELT

Massive amounts of healthcare [3], [4] and disaster management data [48], [49] are generated by many different organizations. The ETL system is unable to manage the order data of terabytes and petabytes because its operation is typically supported by only one machine known as the ETL server. However, the complex nature of big data can potentially be managed by paralleling or distributing the data processing in clusters. In relation to this, the DSS community suggested that the ETL process should be divided amongst a number of cluster nodes to solve the issue. Thus, ETL can perform better due to the parallel management of source data partitions by different components of the ETL process [50]. One option in such cases is the use of Hadoop, which involves ELT or 'extraction' of data from sources, 'loading' data in the HBase database, and 'transformation' and integration of data in the targeted form in the Hive [51].

3) SPATIAL ETL

Integrating GIS technologies with data warehousing ability makes it possible to obtain SDWs. ETL was broadened to SETL in [38] for this purpose. Meanwhile, the extra plug-in

TABLE 2. Summary of data transformation studies related to big data.

Author(s)	Year	Description	Advantages	Drawbacks	Database Model	Big Data Category
Kathiravelu et al. [14]	2019	An on-demand acquisition and storage of data approach was proposed according to analysis requirements. In this way, a major limitation of eager ETL is solved because loading the entire data source content in an integrated data repository is an unnecessary preliminary procedure.	- Supports the hybrid method of eager–lazy ETL - Requires human intervention in the ETL loop - Allows virtual sharing	- Only applied and validated on medical data - Unable to support geospatial data - Unable to support virtual DWs	RDBMS NoSQL	Volume Variety
Homayouni et al. [63]	2018	The approach in this study defined certain rules to ensure the quality of source and target data and automatically builds mapping relations, which can support 1-1, M-1, M-M mapping, between entities. The current approach validated data completeness and consistency using current mapping.	- Develops a balancing test approach - Automatically builds mapping relations between entities	- Fails to address limitations of the approach - Evaluation in other domains is required - Unable to support geospatial data	RDBMS	Volume Veracity
Lupa et al. [40]	2018	The method in this study was developed as a plug-in 'Spatial extension for Talend' to enable complete data integration of features amongst the Database of Topographic Objects 500 and 10k scale databases by applying Talend Open Studio for data integration to formulate a SETL process.	- Extraction, transformation and loading of a range of data formats are facilitated by multiple constituents of Talend. - Supports geospatial data	- Needs human intervention - Unable to support different scales	RDBMS Shapefiles	
Baldacci et al. [9]	2017	This approach took the form of query–extract-transform-load (QETL) process to supply a multidimensional cube, and involved loading facts acquired from source data, which were exclusively provided to the cube when required to address an online analytical processing query and set down facts when space must be freed for loading different facts.	- The optimization procedure is a major feature of QETL that involves the use of specific data provider characteristics to acquire the necessary data efficiently	- Has only been validated in a single scenario - Unable to support geospatial data - Runs only on a single machine	RDBMS	Volume
Bala et al. [50]	2017	P-ETL, a novel approach containing five procedures, was proposed as an alternative to the three traditional ETL steps. This method outperforms traditional ETLs in big data scenarios.	- Applicable to parallelized data processing method - Adds partitioning to and reduces steps in the traditional ETL process	- Must be validated by a commonly used benchmark - Unable to support geospatial data	RDBMS NoSQL	Volume

'Spatial extension for Talend' was developed in [40] by applying Talend Open Studio for data integration to formulate an SETL process and enable complete data integration of geospatial features amongst databases. The extraction, transformation and loading of a range of data formats are facilitated by the multiple constituents of Talend. However, shapefiles with geometrical data are unreadable using a conventional Talend application. Jo and Lee [52] proposed D_ETL as delayed ELT steps by applying the parallel data processing method to deal with massive amounts of IoT geospatial data. The experimental results showed a better performance of the D_ETL data preparation compared with traditional ETL and ELT when dealing with large amounts of data. However, D_ETL demonstrated poor performance in simple data analysis and showed inability to support geospatial big data queries.

SpatialHadoop [47], GeoSpark [30] and other geospatial systems have been proposed to address the issues related to geospatial big data processing and analysis. However, the mechanism underlying the majority of such systems involves the introduction of types or functions of geospatial data into current big data systems, thus not providing straightforward implementations.

C. INTEGRATION METHODS

Domain requirements play a key role in selecting a suitable data integration method or technique. The classification of integration methods into the two main types of database data models and semantic integration is proposed to address the large quantity of different BDI methods and techniques.

1) DATABASE DATA MODELS

a: RDBMS

The widely used relational data model is the basis of relational database management systems (RDBMS). This technology is incorporated in classical DWs and legacy systems [53]. Apart from facilitating the storage and processing of small and structured data, RDBMS is also the preferred technology for GIS data storage. PostgreSQL and the related PostGIS geospatial extension demonstrate advantages, such as efficient functions in vector and raster models [54]. A wide range of models, schemas or formats is used by various organizations. Data heterogeneity can be (1) syntactic or (2) schematic [55]. Syntactic heterogeneity is caused by the usage of different database systems, such as relational or object-oriented databases and geometric representations (e.g. raster or vector representations). Schematic heterogeneity is caused by the use of dissimilar data models for representing identical actual objects. RDBMS ensures that data are consistent and intact in the context of data management. However, certain problems persist in the storage, access and maintenance of large volumes of data, management of semi-structured and unstructured data, and achievement of horizontal scalability [56].

b: NoSQL

RDBMS reaches the maximum capacity and is insufficient in managing big data [57]. Thus, NoSQL databases are used for such issues. At present, massive amounts of unstructured data are available in GIS. Previous studies explored the NoSQL database to store and query geospatial data. Vathy-Fogarassy and Hugyák [7] proposed a data integration framework that supports different data models on GIS. This framework allows the retrieval of data from RDBMS and NoSQL databases at the same time, integrates different data sources, and considers causal users. NoSQL implements various methods and workflows without requiring programming expertise from users. Zhang *et al.* [58] suggested an approach for the storage of large geospatial data based on the MongoDB nonrelational database. MongoDB and Python scripting languages have achieved superior work compared with traditional relational databases. Rainho and Bernardino [59] proposed a web GIS using the NoSQL database to increase the efficiency of obtaining a large amount of GIS data from the web. Web GIS adopts MongoDB as NoSQL database and renders geospatial data in the GeoJSON format through a web service. The study outcomes demonstrated that storing and retrieving unstructured geospatial data using MongoDB achieve better results compared with RDBMS databases.

In addition, MongoDB provides competent spatial operators and allows the storage of various data structures. NoSQL systems compensate for the gap caused by storing big data and maintain performance by supporting many data models, such as column stores, key-value stores, document databases and graph databases [60]. NoSQL databases generally have many veracity issues because they fail to apply data integrity or durability techniques unlike relational databases, which apply strict data consistency policies to guarantee the simultaneous delivery of similar data to all users. Khalfi *et al.* [61] presented another technique for the big data environment to address capacity issues and ensure geospatial data consistency in a document-based NoSQL framework through a particular consistency approval workflow. GeoJSON used schema as a logical model to optimize both relational and NoSQL databases along with semantic constraints to support consistent storage.

2) SEMANTIC INTEGRATION

a: Linked data

Semantic web allows universal access to interlinked web data by using the linked data paradigm. Data are supplied in standardized formats and connected to additional web data sources with constant expansion in the volume of geospatial data; such geospatial data can be accessed through the web in state-managed SDIs and activities of a voluntary, scientific and corporate nature, and influenced by factors, such as legislation, market appeal and social interaction [62].

Data can be organized, disseminated, discovered, accessed and integrated in new ways via linked data. Geospatial resources in SDIs can be efficiently identified and shared by exploiting the strengths of linked data, such as the common data model, a mechanism of standardized data access and link-based identification of data. OGC standard-abiding and other geospatial web services demonstrating interoperability have been created to help implement SDI [10]. Accordingly, user expertise is unnecessary for retrieving, accessing and sharing data in the semantic web and SDI.

Linked data can improve the effectiveness of this entire process, help create a universal infrastructure for data sharing based on the publication of data in a resource description framework (RDF) as a common data model, and use hyperlinks to associate dissimilar data. Purss *et al.* [64] proposed a novel OGC standard in the form of discrete global grid system to offer a homogeneous medium for integration and visualization of vector geometry and raster-based geospatial data sources in a similar manner to the conversion of information in a computer graphics workflow to computer screen pixels. The construct of linked widgets for streaming the data domain is expanded in [65] to facilitate the creation of mashups on the linked widgets platform in real time. The underlying principle of this construct is the representation of the entire data processing workflow and the provision of user assistance in developing semantic data streams and end-user mashups as well as the visualization of processed data flows to obtain knowledge in real time. However, these standards are difficult to apply because of geospatial data geometric joining computations and the requirement of user expertise.

b: ONTOLOGIES

Ontologies convey knowledge as a formal characterization of the target domain. Ontologies are typically used in data integration systems [66] and crucial in data semantics because they explicitly conceptualize a domain in a comprehensible manner for machines. Ontologies facilitate semantic interoperability amongst various web applications and services to address the issue of semantic inhomogeneity [6]. An integrated set of inhomogeneous and dissimilar data sources can be queried by end users via the category of systems called ontology-based data access (OBDA) to reduce the demand for IT assistance [67]. Ontology mapping and hybrid ontology can provide an interpretation of schematic (structural) and semantic interoperability, respectively [68].

Several studies have addressed ontology BDI. For example, Abbes and Gargouri [66] proposed an ontology web language supported by NoSQL databases, MongoDB and modular ontologies, with sources equivalent to big data. It produces local ontologies and then generates a universal ontology by formulating local ones. The storage and processing of heterogeneous data from more than one source in their initial form are permitted by big data configurations. Nadal *et al.* [67] developed a structured ontology supported by an RDF in the form of a BDI ontology that facilitates the modelling and integration of developing data from more than one provider. An integrated set of heterogeneous and dissimilar data sources can be queried by end users via OBDA to reduce the demand for IT assistance. However, these studies are unable to analyze massive amounts of data in real time. Moreover, the schema-less nature of NoSQL databases is considered an obstacle in the ontology integration process, and schema maintenance in response to domain requirements will impact ontology access.

IV. RESEARCH CHALLENGES

The study of geospatial big data is still in the development stage although numerous organizations have applied GIS. Volume, variety, velocity and veracity are key features that define big data and emphasize issues of heterogeneity. However, the demand for data integration by users has led to the emergence of a new research field called BDI, which is limited by key features of big data. This section identifies challenges in the following areas of focus: geospatial big data, big earth data, data warehousing, data transformation, database data models, linked data and ontologies.

A. GEOSPATIAL BIG DATA

Geospatial data are acquired by both public and private organizations. The field of geospatial big data management is underdeveloped [16]. Hence, developing technical and theoretical approaches and providing solutions to critical problems are necessary to understand the core of GIS theory. At present, state-of-the-art techniques, including automation, are used to gather geospatial big data, particularly sensor data, thereby creating novel opportunities and threats. A degree of redundancy also occurs due to the high variety of formats, data providers, sources and usage. A range of obstacles is also associated with the continuity of data updating, data harmonization methods and diverse functionalities that professional-level users require. Consequently, the same geospatial data may be acquired several times and final products may vary despite addressing identical areas [69].

Previous studies focused on the textual database of data integration [70]. At the same time, large volumes of image, audio and sensory data are obtained although these types of data are rarely integrated with textual data into a collective knowledge base. Thus, determining necessary database components is important to facilitate the use, sharing and integration of geospatial data and define standards for organization in a geospatial data setting. In addition, the creation of responsive, accessible and easy-to-use web portals, which users can rely on as decision-making supports, is also crucial.

B. BIG EARTH DATA

From the perspective of big geospatial data development and technical uses, the field of BED is attracting a great deal of attention. The domain of GIS has been progressively permeated by theoretical constructs and empirical techniques accompanying the latest advances in cutting-edge computing technology such as NoSQL databases and cloud computing [16]. BED is a subcategory of big data concerned with EO data such as satellite information, weather data and addressing human activities. It is mainly geared towards facilitating examination and insight into earth-based interactions by analyzing and understanding the available data [71]. A variety of skills and technologies must be combined to undertake the complicated process of analyzing BED, which presents substantial difficulties especially with regard to how to store, process and visualize them. Furthermore, it is necessary to formulate a spatio-temporal data model that effectively supports cloud setting. Relational and non-relational databases and the distributed file system are currently the primary strategies for data storage. Given these circumstances, the strain on the storage system can be alleviated by exploiting the various existing approaches for storing data.

C. DATA WAREHOUSING

The reliance on traditional data-processing techniques is insufficient in the age of big data. The critical obstacle for GIS is related to the integration of large datasets. NoSQL-based DW may offer novel features for data analysis, which would not have been possible under classical DW systems. Also, a considerable promise is associated with distribution and parallelization methods of data and processes. Innovative techniques that can integrate heterogeneous geospatial and non-geospatial data in the context of a unified SDW are urgently required. Constructing virtual distributed DWs is also important because data are partially reproduced and shared between organizations. The creation of a novel DW with NewSQL databases is relational. NewSQL and

NoSQL databases have similar properties, such as distributed architectures and massively intensive parallel processing. Moreover, NewSQL databases are a viable way of managing big data.

Although the DW design has been examined by researchers, DW testing is comparatively underexplored [63]. Data quality testing is utilized to validate data sources prior to loading to the target DW. Similarly, techniques have been proposed for target data validation in isolation although these approaches overlook data changes or losses arising from the ETL process.

D. DATA TRANSFORMATION

The ETL process can consolidate data from heterogeneous sources. However, ETL should be re-examined to address the complex nature of big data. Traditional ETL systems typically operate on one machine as the ETL server and are unable to handle large datasets. Therefore, creating new approaches capable of developing cloud computing, MapReduce and NoSQL data models as well as optimizing and enhancing the efficacy of current ETL approaches are necessary. Notably, ETL frameworks proposed in the literature are incompatible with the automated incorporation of humans in the process. Hence, human-in-the-loop ETL processes are manual, repetitive and time-consuming tasks. Distributed ETL processes can promote integrated data sharing and minimize repeated data loading and integration efforts with a negligible bandwidth overhead. However, research gaps still exist in this topic. Additional indicators should be identified for assessing the system architecture's physical properties and evaluating the effort required to design the ETL phase with respect to a particular domain [19]. Examining problems of extracting, transforming and loading geospatial data in the context of multiple geometric representations is also necessary [13]. At present, ETL processes are incapable of effectively integrating the use of real-time data with archive data, such as information stored in data sources.

E. DATABASE DATA MODELS

The identification of comparable relational databases from numerous expansive databases stored in various database management systems (DBMSs) has become increasingly challenging due to the high diversity in database technologies and sizes [72]. Developing an effective storage solution is necessary to read and write geospatial big data [73] because conventional RDBMSs are limited by the storage and analysis of big data. The massive information volume and RDBMS performance in storage and queries of unstructured geospatial data are major GIS issues. These issues are important because on-demand and real-time queries require effective access to large geospatial data volumes across the Internet [74]. Future investigations should address the matter of parallel process for schema comparisons and the implementation of methods facilitating usage of a greater amount of RAM memory than computers can provide at present. Furthermore, data dictionary and domain are additional metadata dataset characteristics that can be considered [75]. Big data has significantly contributed to the management of unstructured NoSQL databases recently. However, geospatial data specificity is still ignored.

F. LINKED DATA

The integration of heterogeneous web data from multiple sources can be effectively achieved through the strategy of linked data in the context of big data, which involves the initial conversion of heterogeneous geospatial data using current linked geospatial data into an integrated RDF that can be read by machines [76]. However, the processing of web content is generally intended for human users and not machines. Furthermore, accomplishing key goals of linked data, namely, linking and integration, can be challenging and expensive [77]. Different organizations and authorities utilize dissimilar models, schemas or formats. The provision of user-friendly GUI integration tools can facilitate the sharing, interpretation and reuse of knowledge [10]. However, the accuracy and completeness of data are still suboptimal.

G. ONTOLOGIES

The fundamental problem in BDI is the automatic construction of the ontology model and identification of potential semantics that are indirectly accessible from other data sources. The manual generation of ontologies is time-intensive and susceptible to errors. Furthermore, maintaining and updating ontologies is a difficult task. Allowing users to gain an integrated perspective on a dynamic heterogeneous group of data sources is complex and referred to in the literature as the data variety challenge [67]. Therefore, the efficient generation of ontologies is a critical research topic at present. Further investigations can construct novel models for extracting ontologies from other frequently used data sources, such as NoSQL databases. User-friendly visualization is a key goal although linked geospatial resources in the RDF format are able to be processed by computing systems. GUI tools for creating explicit semantic descriptions should also be provided. Overexposure of technical details is a limitation of existing SPARQL implementations [10]. Hence, effective user interfaces are needed for queries rather than SPARQL queries.

V. OPEN RESEARCH ISSUES

Direct human understanding is rapidly exceeded by the scale and complex nature of big data. Therefore, machine support is needed for semantic analysis, organization and interpretation to determine the strategic value of such massive and multisourced data [78]. Many studies are unable to support big geospatial data. Table 3 clearly shows the need to investigate holistic geospatial BDI methods. Fig. 2 presents the different relations between reviewed topics using VOSviewer software. Numerous traditional integration options are inaccessible due to their inability to cope with problems of BDI posed by their volume, velocity, variety and veracity. Automation-based integration methods are necessary to

TABLE 3. Summary of big data integration studies in the literature.

DATA INTEGRATION																							
Ref.	Year	Data warehousing									Data transformation					Integration methods					
		Classical		Modern				Parallel architectures		Geospatial data		ETL			ELT	SETL	Database data models			Semantic integration		
		RDBMS	ODBMS	Document	Column	Key-value	Graph	Hadoop	Spark	Spatial	Temporal	Eager	Lazy	Query			RDBMS	NoSQL	Hybrid	Linked data	SDI	Ontologies
[42]	2020											√										
[38]	2020																√					
[54]	2020																	√				
[83]	2020																	√				
[68]	2020																					√
[31]	2020									√												
[76]	2019																			√		
[67]	2019																					√
[27]	2019							√														
[77]	2019																			√		
[52]	2019							√							√							
[28]	2019									√	√											
[87]	2019										√											
[80]	2019																					√
[29]	2019									√	√											
[14]	2019											√	√									
[88]	2018									√	√							√				
[44]	2018									√												
[6]	2018																					√
[43]	2018	√																				
[8]	2018	√		√	√		√															
[40]	2018															√						
[60]	2018																	√				
[59]	2018			√														√		√		
[26]	2017						√															
[19]	2017																					√
[24]	2017							√														
[45]	2017									√												
[9]	2017													√								
[50]	2017														√							
[89]	2017															√						
[7]	2017			√															√			
[61]	2017			√														√				
[49]	2017																			√		
[25]	2017			√																		

TABLE 3. *(Continued.)* **Summary of big data integration studies in the literature.**

[15]	2016							√	√											
[32]	2016									√										
[51]	2016													√						
[66]	2016																			√
[47]	2015							√												
[30]	2015								√											
[90]	2015																	√		
[62]	2015																	√	√	
[65]	2015																	√		

replace existing manual methods. A few of these issues are addressed in the following subsections.

A. DATA SOURCES

The continuous proliferation of data sources and volume of existing data in numerous domains makes it difficult to integrate several geospatial datasets into one dataset [79]. The heterogeneity of data sources across government agencies, private organizations, geospatial resolutions, projections and storage formats has led to significant challenges in geospatial BDI. Various geospatial theories and methods can be used to integrate traditional data as well as address geospatial big data [2]. Given the wide range of data providers, technical tools and considerations are insufficient to achieve geospatial data integration from multiple sources. The consideration of institutional, social, legal and policy specifications are also necessary.

B. SEMANTIC INTEGRATION

Semantic heterogeneity occurs when the same real-world object is interpreted differently by various disciplines or user groups [80]. Semantic heterogeneity can also take the shape of naming heterogeneity, whereby different names are given to the same real-world object or different real-world objects have an identical name. Geospatial data sharing is challenging due to its heterogeneous character and results in data duplication issues [55]. Although manual comparison methods can be used for data integration, the process involved requires a considerable amount of effort in the case of massive datasets [68]. Significant advances have been achieved in the past five years, but addressing issues caused by dissimilar conceptualizations and interpretations of geospatial data, exchanging knowledge between different domains and integrating cross-lingual information still require further investigation [81].

C. DATA QUALITY

Data quality can be interpreted depending on the purpose of its use. Organizations, companies and users are responsible for setting their quality requirements. According to [82], quality has several definitions; for example, 'quality is the degree to which a set of inherent characteristics fulfil the requirements; fitness for use; conformance to requirements'. Furthermore, data quality consists of the following dimensions: data consistency, data deduplication, information completeness, data currency and data accuracy [17]. Researchers who use geospatial big data must comprehend how the behavior of the data provider influences the quality of big data, and assess the quality of big data and potential mistakes, such as positioning accuracy, logical consistency and other accuracy-related features of the data, prior to data analysis. Different data sources can be used to improve the reliability of these findings. Visualizing errors and the management of NoSQL database quality are basic challenges. The intrinsic uncertainty in big datasets, such as crowd-sourced datasets, impedes their development and implementation [2], [83].

D. STORAGE GEOSPATIAL BIG DATA

Data storage plays a crucial role in processing and analyzing massive amounts of structured, semi-structured and unstructured data [16]. Significant quantities of data are paired with performance issues in RDBMS for storage and unstructured data, whilst geospatial data queries are fundamental challenges associated with GIS. Given that efficient access to Internet-based geospatial data is critical in performing on-demand and real-time queries, several studies have recently examined the use of the cloud computing paradigm in solving these issues [45]. Driven by the substantial computational and storage-related capacities of cloud computing infrastructures, many studies have applied NoSQL DBMSs, such as MongoDB and HBase. In view of this, a unified metadata format with an effectively formulated data integration framework is urgently needed. Furthermore, studies should focus on deep learning algorithms, particularly the use of semantic matching and unified format conversion of remote sensing metadata [84].

E. PROCESSING GEOSPATIAL BIG DATA

Processing big data obtained from GIS has become increasingly difficult due to data volume [57]. Understanding geospatial data is important because of their positive effect on a range of consequential domains and applications. The infeasibility of waiting until a complete dataset is obtained is a critical problem of geospatial algorithms in real-time big data processing. Consequently, the distribution and parallelization of geospatial algorithms are essential. Identifying

viable strategies for rapidly processing user requests, such as responses in less than one second, is a main obstacle. Accordingly, adding novel functionalities that promote development in geospatial data frameworks are required. Scientists currently operating in this area can work on frameworks such as GeoSpark [30].

F. BIG EARTH DATA ANALYTICS

The analytical lifecycle of preparation, analysis, mining, and visualization of ample volumes of various types of spatio-temporal data constitutes the analytics of BED. Through this process, comprehension of the earth system can be improved and issues caused by transformations at global and regional level can be better addressed by uncovering a range of relevant information, including patterns, causations, and knowledge [85]. Although there have been several endeavors to create an integrated model for analyzing big data [71], this process of analysis remains a complicated task that calls for the fusion of numerous different skills and technologies.

Undertaking such meta-analysis is frequently made more difficult by the fact that the greatest proportion of data lack structure. In addition, scalability and parallelism are also issues, as with big data in general. Therefore, within this age of big data, geospatial analytics necessitate adaptable architectures capable of employing the existing data as much as possible, effective scaling with data volume, demonstrating compatibility with different modelling frameworks, and offering users options for exploring and visualizing data interactively [86].

G. BIG SPATIOTEMPORAL DATA ANALYTICS

Big spatio-temporal data analytics is necessary to investigate and apply relevant algorithms, frameworks, and solutions for big data generated with space and time stamps [91]. The processes of observation and documentation of both natural and social phenomena have been significantly enhanced by sensing technology innovations as case in point tracking of the COVID-19 crises [4]. However, spatio-temporal analysis has not yet achieved maturity, with many issues still to be overcome, including the pattern types that can be derived from time series data and the applicable techniques and algorithms. The formulation of new techniques of real-time event discovery could help to resolve such issues. Furthermore, it is essential to achieve improved data integration for inclusive and multi-dimensional event identification by taking advantage of the fast proliferation of spatio-temporal data sources. This can have positive implications not only for how events are comprehended scientifically, but also for the operational processes underpinning decision-making associated with events [92].

VI. EMERGING BIG GEOSPATIAL DATA TRENDS

In the present age of big geospatial and EO data, development over last years shows its evolution from conventional areas to more evolutionary application areas such as healthcare risk management, environmental disasters mitigation and prediction, addressing human activities and self-driving vehicles. To address these needs, emerging technologies have been designed and developed. This section identifies big geospatial data trends in the following areas of focus: big geospatial cloud computing, big geospatial data in the context of artificial intelligence (AI) and machine learning (ML), smart geospatial data discovery and geospatial data content understanding.

A. BIG GEOSPATIAL DATA CLOUD COMPUTING

Cloud computing can facilitate computer resources dissemination by ensuring that they are used as effectively as possible with regard to CPU, RAM, network and storage [93]. The field of Computer Sciences is currently undergoing a paradigm change in the direction of cloud computing [94], which has been useful for a number of sophisticated applications and has markedly improved storage and computational cost-effectiveness [95]. Given the open availability of large volumes of geospatial data, appropriate storage, processing, transmission, and analysis of such data present difficulties for conventional SDI. This has prompted the creation of new cloud computing-based technologies, such as GeoRocket, which is among the first cloud-based technologies intended exclusively for geospatial data management [94]. Other technologies running in the cloud and facilitating geospatial datasets to be scientifically analyzed and visualized on a large scale are the Google Earth Engine and the System for EO Data Access, Processing and Analysis for Land Monitoring [37]. However, there is one significant problem associated with cloud computing platforms, namely, vendor lock in. The migration of data can be disrupted by the fact that management and processing functions are dissimilar between cloud platforms [71]. Additionally, the storage and processing of ample geospatial data within remote cloud servers can be subject to delays and energy use due to the geographically specific nature of geospatial data [96].

B. BIG GEOSPATIAL DATA IN THE CONTEXT OF AI AND ML

As computing power, learning algorithms and application scenarios become more sophisticated and diverse, the application of AI in a range of fields has intensified. Geospatial information science benefits significantly from AI, particularly when used alongside big data analysis [97]. Furthermore, BSD analytics can be refined by drawing of new methods, such as explainable AI and interpretable ML [98]. Given the variety of domains in which big geospatial data are relevant (e.g. infection tracking, climate change simulations, disaster management, etc.), research has been focused on supplying geospatial extensions to current ML solutions or formulating entirely new solutions to enable effective analysis and intelligence for existing applications [99]. Nevertheless, further research is needed to determine which geospatial applications are most influential as well as to integrate geospatial techniques and parallelization in this age of big data [100]. The lack of big data homogeneity poses

a particular research difficulty to formulate more advanced algorithms to be used more effectively [101].

C. SMART GEOSPATIAL DATA DISCOVERY

The vast daily production of data poses great difficulties to the field of Earth Sciences in terms of geospatial data discovery and accessibility [102]. In particular, the application of linked data and precise data discovery are affected by the lack of semantic homogeneity of geospatial data [103]. In this context, the creation of geospatial data portals has been proposed as a solution to make big geospatial data more accessible [104]. Jiang *et al.* [102] devised an intelligent system of geospatial data discovery based on the web, whereby metadata user behavior is exploited to mine and use data relevancy. Besides, in the context of intelligent sustainable urban centers and the built environment, quantitative and semantics analysis can benefit from the integration of building information modeling (BIM) and GIS, which can also provide visualization opportunities for knowledge discovery and informed decision-making [105]. Furthermore, building knowledge graphs from multiple sources can be semantically linked in a spatio-temporal manner, thus affording researchers more reliable and effective services [106].

D. GEOSPATIAL DATA CONTENT UNDERSTANDING

Integrating heterogeneous data leads to better data representation and understanding. Conventional survey data can be exploited for research purposes to the greatest degree possible through the fusion of geospatial analytics and big data techniques. For instance, to shed light on how income inequality and health were correlated. Haithcoat *et al.* [107] integrated geospatial big analytics and conventional large-scale survey data. Another important use of big geospatial data is in controlling self-driving vehicles as a new frontier of smart transport, taking advantage of the ability of these types of vehicles to sense the environment and operate with minimal or without human intervention [108]. Furthermore, important insight into travel behavior, traffic flow, and surrounding environment can be gained from knowledge of big geospatial data. Novel systems and tools are needed to successfully explore the existing data archives, given the continuous expansion of EO data [101]. However, data accuracy should be taken into account to gain an inclusive comprehension of such data [109].

VII. CONCLUSION

We provide a review of geospatial BDI methods and infrastructures. The classification of BDI approaches, including data warehousing, data transformation and integration methods, is proposed in this study. A large number of studies related to data warehousing and ETL tools is explored and summarized, and their drawbacks and limitations are highlighted. Many studies focused on and were limited to the so-called structured data whilst ignoring unstructured data, especially geospatial big data. Our study clearly shows that holistic geospatial BDI methods need further investigation.

Lastly, several persistent challenges, research issues and trends have been discussed on the basis of the current big data era. This review covered data sources, semantic integration, data quality, processing and storage of geospatial data, and improvements in these research insights will be beneficial to this field.

REFERENCES

[1] H. Goyal, C. Sharma, and N. Joshi, "An integrated approach of GIS and spatial data mining in big data," *Int. J. Comput. Appl.*, vol. 169, no. 11, pp. 8887–8975, 2017.

[2] S. Li, S. Dragicevic, F. A. Castro, M. Sester, S. Winter, A. Coltekin, C. Pettit, B. Jiang, J. Haworth, A. Stein, and T. Cheng, "Geospatial big data handling theory and methods: A review and research challenges," *ISPRS J. Photogramm. Remote Sens.*, vol. 115, pp. 119–133, May 2016.

[3] J. P. Mcglothlin, A. Madugula, and I. Stojic, "The virtual enterprise data warehouse for healthcare," in *Proc. 10th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, vol. 5, pp. 469–476, Feb. 2017.

[4] C. Zhou *et al.*, "COVID-19: Challenges to GIS with big data," *Geography Sustainability*, vol. 1, no. 1, pp. 77–87, Mar. 2020.

[5] M. Lenzerini, "Data integration: A theoretical perspective," in *Proc. 21st ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst. (PODS)*, 2002, pp. 233–246.

[6] O. El Hajjamy, L. Alaoui, and M. Bahaj, "Semantic integration of heterogeneous classical data sources in ontological data warehouse," in *Proc. Int. Conf. Learn. Optim. Algorithms, Theory Appl.*, May 2018, pp. 1–8.

[7] Á. Vathy-Fogarassy and T. Hugyák, "Uniform data access platform for SQL and NoSQL database systems," *Inf. Syst.*, vol. 69, pp. 93–105, Sep. 2017.

[8] M. Golfarelli and S. Rizzi, "From star schemas to big data: 20+ years of data warehouse research," in *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*. Cham, Switzerland: Springer, 2018, pp. 93–107.

[9] L. Baldacci, M. Golfarelli, S. Graziani, and S. Rizzi, "QETL: An approach to on-demand ETL from non-owned data sources," *Data Knowl. Eng.*, vol. 112, pp. 17–37, Nov. 2017.

[10] P. Yue, X. Guo, M. Zhang, L. Jiang, and X. Zhai, "Linked data and SDI: The case on Web geoprocessing workflows," *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 245–257, Apr. 2016.

[11] X. L. Dong and D. Srivastava, "Big data integration," in *Proc. IEEE 29th Int. Conf. Data Eng.*, Apr. 2013, pp. 1245–1248.

[12] R. Flowerdew, "Spatial data integration," *Geogr. Inf. Syst.*, vol. 1, pp. 375–387, 1991.

[13] M. Ponjavic, A. Karabegovic, E. Ferhatbegovic, and I. Besic, "Spatial data integration in heterogeneous information systems' environment," in *Proc. 42nd Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2019, pp. 1559–1564.

[14] P. Kathiravelu, A. Sharma, H. Galhardas, P. Van Roy, and L. Veiga, "On-demand big data integration," *Distrib. Parallel Databases*, vol. 37, no. 2, pp. 273–295, Jun. 2019.

[15] A. Eldawy and M. F. Mokbel, "The era of big spatial data: A survey," *Found. Trends Databases*, vol. 6, nos. 3–4, pp. 163–273, 2016.

[16] X. Yao and G. Li, "Big spatial vector data management: A review," *Big Earth Data*, vol. 2, no. 1, pp. 108–129, Jan. 2018.

[17] J. Liu, J. Li, W. Li, and J. Wu, "Rethinking big data: A review on the data quality and usage issues," *ISPRS J. Photogramm. Remote Sens.*, vol. 115, pp. 134–142, May 2016.

[18] (2013). *The Four V's of Big Data*. Accessed: Jan. 7, 2020. [Online]. Available: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[19] F. D. Tria, E. Lefons, and F. Tangorra, "Evaluation of data warehouse design methodologies in the context of big data," in *Proc. Int. Conf. Big Data Anal. Knowl. (DaWaK)*, vol. 10440, Aug. 2017, pp. 3–18.

[20] J.-G. Lee and M. Kang, "Geospatial big data: Challenges and opportunities," *Big Data Res.*, vol. 2, no. 2, pp. 74–81, Jun. 2015.

[21] I. A. Ajah and H. F. Nweke, "Big data and business analytics: Trends, platforms, success factors and applications," *Big Data Cognit. Comput.*, vol. 3, no. 2, p. 32, Jun. 2019.

[22] N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," *Scientometrics*, vol. 84, no. 2, pp. 523–538, Aug. 2010.

[23] R. Venkatraman and S. Venkatraman, "Big data infrastructure, data visualisation and challenges," in *Proc. 3rd Int. Conf. Big Data Internet Things (BDIOT)*, 2019, pp. 13–17.
[24] L. W. Santoso, "Data warehouse with big data technology for higher education," *Procedia Comput. Sci.*, vol. 124, pp. 93–99, Dec. 2017.
[25] Z. Bicevska and I. Oditis, "Towards NoSQL-based data warehouse solutions," *Procedia Comput. Sci.*, vol. 104, pp. 104–111, Jan. 2017.
[26] H. Akid and M. Ben Ayed, "Towards NoSQL graph data warehouse for big social data analysis," in *Proc. Int. Conf. Intell. Syst. Design Appl.*, 2017, pp. 965–973.
[27] Z. Han, F. Qin, C. Cui, Y. Liu, L. Wang, and P. Fu, "Mr4Soil: A MapReduce-based framework integrated with GIS for soil erosion modelling," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 3, p. 103, Feb. 2019.
[28] D. Rammer, S. L. Pallickara, and S. Pallickara, "ATLAS: A distributed file system for spatiotemporal data," in *Proc. 12th IEEE/ACM Int. Conf. Utility Cloud Comput.*, Dec. 2019, pp. 11–20.
[29] S. Wang, Y. Zhong, and E. Wang, "An integrated GIS platform architecture for spatiotemporal big data," *Future Gener. Comput. Syst.*, vol. 94, pp. 160–172, May 2019.
[30] J. Yu, J. Wu, and M. Sarwat, "GeoSpark: A cluster computing framework for processing large-scale spatial data," in *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2015, pp. 4–7.
[31] H. Haroun, A. R. Ghomari, M. Lahlouh, and A. Mehdi, "Towards a spatial data warehouse for occupational health risk management," in *Proc. 1st Int. Conf. Innov. Res. Appl. Sci., Eng. Technol. (IRASET)*, Apr. 2020, pp. 1–6.
[32] H. Baazaoui-Zghal, "Fuzzy ontology-based spatial data warehouse for context-aware search and recommendation," in *Proc. 11th Int. Joint Conf. Softw. Technol.*, 2016, pp. 161–166.
[33] Q. Li, S. J. Yang, H. J. Huang, and Y. H. Zhou, "Geo-spatial big data storage based on NoSQL database," *Geomatics Inf. Sci. Wuhan Univ*, vol. 42, no. 2, pp. 163–169, 2017.
[34] P. Baumann, D. Misev, V. Merticariu, and B. P. Huu, "Datacubes: Towards space/time analysis-ready data," in *Service-Oriented Mapping*. Cham, Switzerland: Springer, 2019, pp. 269–299.
[35] M. Sudmanns, D. Tiede, S. Lang, H. Bergstedt, G. Trost, H. Augustin, A. Baraldi, and T. Blaschke, "Big Earth data: Disruptive changes in Earth observation data management and analysis?" *Int. J. Digit. Earth*, vol. 13, no. 7, pp. 832–850, Jul. 2020.
[36] G. A. Pagani and L. Trani, "Data cube and cloud resources as platform for seamless geospatial computation," in *Proc. 15th ACM Int. Conf. Comput. Frontiers*, May 2018, pp. 293–298.
[37] V. C. F. Gomes, G. R. Queiroz, and K. R. Ferreira, "An overview of platforms for big Earth observation data management and analysis," *Remote Sens.*, vol. 12, no. 8, pp. 1–25, 2020.
[38] U. Drescek, M. K. Fras, J. Tekavec, and A. Lisec, "Spatial ETL for 3D building modelling based on unmanned aerial vehicle data in semi-urban areas," *Remote Sens.*, vol. 12, no. 12, p. 1972, Jun. 2020.
[39] S. Laraichi, A. Hammani, and A. Bouignane, "Data integration as the key to building a decision support system for groundwater management: Case of Saiss aquifers, Morocco," *Groundwater Sustain. Develop.*, vols. 2–3, pp. 7–15, Aug. 2016.
[40] M. Lupa, W. Sarlej, and K. Adamek, "Harmonization of datasets in the frame of spatial data infrastructure using ETL tools: A case study of BDOT500 and BDOT10k databases," in *Proc. Baltic Geodetic Congr. (BGC Geomatics)*, Jun. 2018, pp. 217–220.
[41] N. Biswas, A. Sarkar, and K. C. Mondal, "Efficient incremental loading in ETL processing for real-time data integration," *Innov. Syst. Softw. Eng.*, vol. 16, no. 1, pp. 53–61, Mar. 2020.
[42] J. Sreemathy, I. J. V, S. Nisha, C. P. I, and G. P. R. M., "Data integration in ETL using TALEND," in *Proc. 6th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2020, pp. 1444–1448.
[43] H. Moulai and H. Drias, "From data warehouse to information warehouse: Application to social media," in *Proc. Int. Conf. Learn. Optim. Algorithms, Theory Appl.*, 2018, pp. 1–6.
[44] M. Mazzei and S. Di Guida, "Spatial data warehouse and spatial OLAP in indoor/outdoor cultural environments," in *Proc. Int. Conf. Comput. Sci. Appl.* Cham, Switzerland: Springer, May 2018, pp. 233–250.
[45] S. Bimonte, M. Zaamoune, and P. Beaune, "Conceptual design and implementation of spatial data warehouses integrating regular grids of points," *Int. J. Digit. Earth*, vol. 10, no. 9, pp. 901–922, Sep. 2017.
[46] M. Barkhordari and M. Niamanesh, "Atrak: A MapReduce-based data warehouse for big data," *J. Supercomput.*, vol. 73, no. 10, pp. 4596–4610, Oct. 2017.
[47] A. Eldawy and M. F. Mokbel, "SpatialHadoop: A MapReduce framework for spatial data," in *Proc. IEEE 31st Int. Conf. Data Eng.*, Apr. 2015, pp. 1352–1363.
[48] A. G. Rumson, S. H. Hallett, and T. R. Brewer, "Coastal risk adaptation: The potential role of accessible geospatial big data," *Mar. Policy*, vol. 83, pp. 100–110, Sep. 2017.
[49] V. Bhanumurthy, K. R. M. Rao, G. J. Sankar, and P. V. Nagamani, "Spatial data integration for disaster/emergency management: An Indian experience," *Spatial Inf. Res.*, vol. 25, no. 2, pp. 303–314, Apr. 2017.
[50] M. Bala, O. Boussaid, and Z. Alimazighi, "A fine-grained distribution approach for ETL processes in big data environments," *Data Knowl. Eng.*, vol. 111, pp. 114–136, Sep. 2017.
[51] H.-K. Lin, J. A. Harding, and C.-I. Chen, "A hyperconnected manufacturing collaboration system using the semantic Web and Hadoop ecosystem system," *Procedia CIRP*, vol. 52, pp. 18–23, Jan. 2016.
[52] J. Jo and K.-W. Lee, "MapReduce-based D_ELT framework to address the challenges of geospatial big data," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 11, p. 475, Oct. 2019.
[53] H. Dhayne, R. Haque, R. Kilany, and Y. Taher, "In search of big medical data integration solutions—A comprehensive survey," *IEEE Access*, vol. 7, pp. 91265–91290, 2019.
[54] D. Guo and E. Onstein, "State-of-the-art geospatial information processing in NoSQL databases," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 5, p. 331, May 2020.
[55] F. Yu, D. A. McMeekin, L. Arnold, and G. West, "Semantic Web technologies automate geospatial data conflation: Conflating points of interest data for emergency response services," in *Proc. 14th Int. Conf. Location Based Services (LBS)*. Cham, Switzerland: Springer, 2018, pp. 111–131.
[56] F. Gao, P. Yue, Z. Wu, and M. Zhang, "Geospatial data storage based on HBase and MapReduce," in *Proc. 6th Int. Conf. Agro-Geoinformat.*, Aug. 2017, pp. 1–4.
[57] D. R. D. Almeida, C. D. S. Baptista, F. G. D. Andrade, and A. Soares, "A survey on big data for trajectory analytics," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 2, pp. 1–24, 2020.
[58] X. Zhang, W. Song, and L. Liu, "An implementation approach to store GIS spatial data on NoSQL database," in *Proc. 22nd Int. Conf. Geoinformat.*, Jun. 2014, pp. 4–8.
[59] F. D. C. Rainho and J. Bernardino, "Web GIS: A new system to store spatial data using GeoJSON in MongoDB," in *Proc. 13th Iberian Conf. Inf. Syst. Technol. (CISTI)*, Jun. 2018, pp. 1–6.
[60] M. R. Ahmed, M. R. Ahmed, M. A. Khatun, M. A. Ali, and K. Sundaraj, "A literature review on NoSQL database for big data processing," *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 902–906, 2018.
[61] B. Khalfi, C. De Runz, S. Faiz, and A. Herman, "A new methodology for storing consistent fuzzy geospatial data in big data environment," *IEEE Trans. Big Data*, Jul. 2017.
[62] S. Wiemann and L. Bernard, "Spatial data fusion in spatial data infrastructures using linked data," *Int. J. Geogr. Inf. Sci.*, vol. 30, no. 4, pp. 613–636, 2016.
[63] H. Homayouni, S. Ghosh, and I. Ray, "An approach for testing the extract-transform-load process in data warehouse systems," in *Proc. 22nd Int. Database Eng. Appl. Symp. (IDEAS)*, 2018, pp. 236–245.
[64] M. B. J. Purss, R. Gibb, F. Samavati, P. Peterson, and J. Ben, "The OGC discrete global grid system core standard: A framework for rapid geospatial integration," in *Proc. Int. Geosci. Remote Sens. Symp.*, Nov. 2016, pp. 3610–3613.
[65] A. M. Tjoa, P. Wetz, E. Kiesling, T.-D. Trinh, and B.-L. Do, "Integrating streaming data into semantic mashups," *Procedia Comput. Sci.*, vol. 72, pp. 1–4, 2015.
[66] H. Abbes and F. Gargouri, "Big data integration: A MongoDB database and modular ontologies based approach," *Procedia Comput. Sci.*, vol. 96, pp. 446–455, Jan. 2016.
[67] S. Nadal, O. Romero, A. Abelló, P. Vassiliadis, and S. Vansummeren, "An integration-oriented ontology to govern evolution in big data ecosystems," *Inf. Syst.*, vol. 79, pp. 3–19, Jan. 2019.
[68] C. Prudhomme, T. Homburg, J.-J. Ponciano, F. Boochs, C. Cruz, and A.-M. Roxin, "Interpretation and automatic integration of geospatial data into the semantic Web: Towards a process of automatic geospatial data interpretation, classification and integration using semantic technologies," *Computing*, vol. 102, no. 2, pp. 365–391, Feb. 2020.
[69] Z. Li, "Geospatial big data handling with high performance computing: Current approaches and future directions," Jul. 2019, *arXiv:1907.12182*. [Online]. Available: http://arxiv.org/abs/1907.12182

[70] X. L. Dong and T. Rekatsinas, "Data integration and machine learning: A natural synergy," in *Proc. Int. Conf. Manage. Data*, May 2018, pp. 1645–1650.
[71] P. Merritt, H. Bi, B. Davis, C. Windmill, and Y. Xue, "Big Earth data: A comprehensive analysis of visualization analytics issues," *Big Earth Data*, vol. 2, no. 4, pp. 321–350, Oct. 2018.
[72] D. G. D. Reis, M. Ladeira, M. Holanda, and M. de Carvalho Victorino, "Large database schema matching using data mining techniques," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2018, pp. 523–530.
[73] L. Zhang, Q. Li, Y. Li, and Y. Cai, "A distributed storage model for healthcare big data designed on HBase," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 4101–4105.
[74] L. van den Brink, P. Barnaghi, J. Tandy, G. Atemezing, R. Atkinson, B. Cochrane, Y. Fathy, R. García Castro, A. Haller, A. Harth, K. Janowicz, A. Kolozali, B. van Leeuwen, M. Lefrançois, J. Lieberman, A. Perego, D. Le-Phuoc, B. Roberts, K. Taylor, and R. Troncy, "Best practices for publishing, retrieving, and using spatial data on the Web," *Semantic Web*, vol. 10, no. 1, pp. 95–114, Dec. 2018.
[75] A. Holemans, J.-P. Kasprzyk, and J.-P. Donnay, "Coupling an unstructured NoSQL database with a geographic information system," in *Proc. 10th Int. Conf. Adv. Geogr. Inf. Syst. Appl. Serv. Coupling*, 2018, pp. 23–28.
[76] S. Athanasiou, G. Giannopoulos, D. Graux, N. Karagiannakis, J. Lehmann, A. C. Ngomo, K. Patroumpas, M. A. Sherif, and D. Skoutas, "Big POI data integration with linked data technologies," *Adv. Database Technol.-EDBT*, vol. 2019, pp. 477–488, Mar. 2019.
[77] M. Mountantonakis and Y. Tzitzikas, "Large-scale semantic integration of linked data: A survey," *ACM Comput. Surv.*, vol. 52, no. 5, pp. 1–40, Oct. 2019.
[78] D. J. Kim, J. Hebeler, V. Yoon, and F. Davis, "Exploring determinants of semantic Web technology adoption from IT professionals' perspective: Industry competition, organization innovativeness, and data management capability," *Comput. Hum. Behav.*, vol. 86, pp. 18–33, Sep. 2018.
[79] H. Abbes and F. Gargouri, "MongoDB-based modular ontology building for big data integration," *J. Data Semantics*, vol. 7, no. 1, pp. 1–27, Mar. 2018.
[80] L. Ding, G. Xiao, D. Calvanese, and L. Meng, "Consistency assessment for open geodata integration: An ontology-based approach," *Geoinformatica*, pp. 1–26, Dec. 2019.
[81] M. Kokla and E. Guilbert, "A review of geospatial semantic information modeling and elicitation approaches," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 3, p. 31, 2020.
[82] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Sci. J.*, vol. 14, p. 2, May 2015.
[83] A. A. Frozza and R. D. S. Mello, "JS4Geo: A canonical JSON schema for geographic data suitable to NoSQL databases," *GeoInformatica*, vol. 24, no. 4, pp. 987–1019, Oct. 2020.
[84] J. Fan, J. Yan, Y. Ma, and L. Wang, "Big data integration in remote sensing across a distributed metadata-based spatial infrastructure," *Remote Sens.*, vol. 10, no. 1, pp. 1–20, 2018.
[85] C. Yang, M. Yu, Y. Li, F. Hu, Y. Jiang, Q. Liu, D. Sha, M. Xu, and J. Gu, "Big Earth data analytics: A survey," *Big Earth Data*, vol. 3, no. 2, pp. 83–107, Apr. 2019.
[86] C. Robertson, C. Chaudhuri, M. Hojati, and S. A. Roberts, "An integrated environmental analytics system (IDEAS) based on a DGGS," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 214–228, Apr. 2020.
[87] K. Graff, C. Lissak, Y. Thiery, O. Maquaire, S. Costa, M. Medjkane, and B. Laignel, "Characterization of elements at risk in the multirisk coastal context and at different spatial scales: Multi-database integration (Normandy, France)," *Appl. Geography*, vol. 111, Oct. 2019, Art. no. 102076.
[88] K. Bin and T. Rahim, "Spatiotemporal applications of big data," *Int. J. Comput. Appl.*, vol. 181, no. 21, pp. 5–10, Oct. 2018.
[89] R. P. D. Nath, K. Hose, T. B. Pedersen, and O. Romero, "SETL: A programmable semantic extract-transform-load framework for semantic data warehouses," *Inf. Syst.*, vol. 68, pp. 17–43, Aug. 2017.
[90] W. Li, M. Song, B. Zhou, K. Cao, and S. Gao, "Performance improvement techniques for geospatial Web services in a cyberinfrastructure environment—A case study with a disaster management portal," *Comput., Environ. Urban Syst.*, vol. 54, pp. 314–325, Nov. 2015.
[91] C. Yang, K. Clarke, S. Shekhar, and C. V. Tao, "Big spatiotemporal data analytics: A research and innovation frontier," *Int. J. Geographical Inf. Sci.*, vol. 34, no. 6, pp. 1075–1088, Jun. 2020.
[92] M. Yu *et al.*, "Spatiotemporal event detection: A review," *Int. J. Digit. Earth*, pp. 1–27, Mar. 2020.
[93] Y. Li, M. Yu, M. Xu, J. Jhang, D. Sha, Q. Liu, and C. Yang, "Big data and cloud computing," *Manual Digit. Earth*, pp. 325–355, Nov. 2019.
[94] M. Krämer, "GeoRocket: A scalable and cloud-based data store for big geospatial files," *SoftwareX*, vol. 11, Jan. 2020, Art. no. 100409.
[95] S. Bebortta, S. K. Das, M. Kandpal, R. K. Barik, and H. Dubey, "Geospatial serverless computing: Architectures, tools and future directions," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 5, pp. 1–26, 2020.
[96] J. Das, A. Mukherjee, S. K. Ghosh, and R. Buyya, "Spatio-fog: A green and timeliness-oriented fog computing model for geospatial query resolution," *Simul. Model. Pract. Theory*, vol. 100, Apr. 2020, Art. no. 102043.
[97] D. Li, Z. Shao, and R. Zhang, "Advances of geo-spatial intelligence at LIESMARS," *Geo-spatial Inf. Sci.*, vol. 23, no. 1, pp. 40–51, Jan. 2020.
[98] B. Huang and J. Wang, "Big spatial data for urban and environmental sustainability," *Geo-Spatial Inf. Sci.*, vol. 23, no. 2, pp. 125–140, Apr. 2020.
[99] I. Sabek and M. F. Mokbel, "Machine learning meets big spatial data," in *Proc. IEEE 36th Int. Conf. Data Eng. (ICDE)*, Apr. 2020, pp. 1782–1785.
[100] Z. Li, W. Tang, Q. Huang, E. Shook, and Q. Guan, "Introduction to big data computing for geospatial applications," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 8, p. 487, Aug. 2020.
[101] K. Alonso, D. Espinoza-Molina, and M. Datcu, "Multilayer architecture for heterogeneous geospatial data analytics: Querying and understanding EO archives," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 3, pp. 791–806, Mar. 2017.
[102] Y. Jiang, Y. Li, C. Yang, F. Hu, E. Armstrong, T. Huang, D. Moroni, L. McGibbney, F. Greguska, and C. Finch, "A smart Web-based geospatial data discovery system with oceanographic data as an example," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 2, p. 62, Feb. 2018.
[103] K. Sun, Y. Zhu, P. Pan, K. Luo, D. Wang, and Z. Hou, "Morphology-ontology of geospatial data and its application in data discovery," in *Proc. 23rd Int. Conf. Geoinformat.*, Jun. 2015, pp. 1–6.
[104] Y. Jiang, Y. Li, C. Yang, F. Hu, E. M. Armstrong, T. Huang, D. Moroni, L. J. McGibbney, and C. J. Finch, "Towards intelligent geospatial data discovery: A machine learning framework for search ranking," *Int. J. Digit. Earth*, vol. 11, no. 9, pp. 956–971, Sep. 2018.
[105] M. Breunig, P. E. Bradley, M. Jahn, P. Kuper, N. Mazroob, N. Rösch, M. Al-Doori, E. Stefanakis, and M. Jadidi, "Geospatial data management research: Progress and future directions," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 2, p. 95, Feb. 2020.
[106] Y. Zhu, "Geospatial semantics, ontology and knowledge graphs for big Earth data," *Big Earth Data*, vol. 3, no. 3, pp. 187–190, Jul. 2019.
[107] T. L. Haithcoat, E. E. Avery, K. A. Bowers, R. D. Hammer, and C.-R. Shyu, "Income inequality and health: Expanding our understanding of state-level effects by using a geospatial big data approach," *Social Sci. Comput. Rev.*, pp. 1–19, Sep. 2019.
[108] S. Shang, J. Shen, J.-R. Wen, and P. Kalnis, "Deep understanding of big geospatial data for self-driving cars," *Neurocomputing*, pp. 1–2, 2020.
[109] S. Zhang, B. Zhao, Y. Tian, and S. Chen, "Stand with# StandingRock: Envisioning an epistemological shift in understanding geospatial big data in the 'post-truth' era," *Ann. Amer. Assoc. Geogr.*, pp. 1–21, Aug. 2020.

SOHAIB AL-YADUMI received the bachelor's degree in computer science from the University of Science and Technology, Yemen, in 2003, and the master's degree from the Arab Academy for Management, Banking and Financial Sciences, Yemen, in 2007. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Taylor's University, Malaysia. His research interests include spatial database and geographic information systems (GIS), and big data integration and analytics.

TAN EE XION received the B.Sc. degree in computer science from Monash University, Australia, in 2001, and the M.IT. (minor thesis) and Ph.D. degrees from Monash University Malaysia, in 2005 and 2016, respectively. She is currently a Senior Lecturer with the Life Sciences (Health Informatics and Analytics), School of Pharmacy, International Medical University (IMU), Malaysia. Her research interests include information systems, e-learning teaching and learning assessment, social media, big data, computational linguistics, and cognitive science research.

SHARON GOH WEI WEI received the Ph.D. degree major in E-learning from the University of Derby, U.K. She has been in the academic field for the past 20 years. She has been working on learning, teaching and assessment design strategies, online learning technologies, social media technologies, machine learning, and the Internet of Things research throughout these 16 years. She is currently a Senior Lecturer with the School of Computer Science and Engineering, Taylor's University. She has experience in both qualitative and quantitative methodologies. She emphasized her familiarity with online technologies due to her experience of using them in teaching as she likes to develop technology-rich classrooms. She has practical skills, hands-on experience, and educational credentials to make a significant difference to the university.

PATRICE BOURSIER received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University Pierre et Marie Curie (now part of Sorbonne Université Paris), France, in 1981.

He received the National Accreditation for Full Professorship from the French Ministry of Higher Education and the University of Paris-South (now part of Paris-Saclay University), France, in 1996. Since 10 February 2020, he has been appointed as a Professor of Computer Science with the International Medical University, in charge of developing new undergraduate and postgraduate programmes related to digital health and health analytics. Prior to this, he had been a Full Professor and the Head of the School of Computing with Taylor's University, from 2018 to 2019, and a Full Professor and the Dean of the STEM Faculty with the International University of Malaya-Wales, from 2014 to 2017. He had previously been a Full Professor of Computer Science with the University of La Rochelle, France, from 1997 to 2014, and an Associate Professor with the University of Paris-South, France, from 1984 to 1997. He has authored and coauthored more than 100 articles in journals, conference proceedings, and book reviews. He has graduated 12 Ph.D. students in France and six Joint Ph.D. students in Malaysia. He is currently supervising two Ph.D. students in Malaysia. His research interests include spatial databases and geographic information systems (GIS), big data, and Artificial Intelligence.