Chapter 9
Big Data and Cloud Computing

Yun Li, Manzhu Yu, Mengchao Xu, Jingchao Yang, Dexuan Sha, Qian Liu and Chaowei Yang

Abstract Big data emerged as a new paradigm to provide unprecedented content and value for Digital Earth. Big Earth data are increasing tremendously with growing heterogeneity, posing grand challenges for the data management lifecycle of storage, processing, analytics, visualization, sharing, and applications. During the same time frame, cloud computing emerged to provide crucial computing support to address these challenges. This chapter introduces Digital Earth data sources, analytical methods, and architecture for data analysis and describes how cloud computing supports big data processing in the context of Digital Earth.

Keywords Geoscience · Spatial data infrastructure · Digital transformation · Big data architecture

9.1 Introduction

Digital Earth refers to the virtual representation of the Earth we live in. It represents the Earth in the digital world from data to model. Data are collected and models are abstracted to build the digital reality. Massive amounts of data are generated from various sensors deployed to observe our home planet while building Digital Earth. The term "big data" was first presented by NASA researchers to describe the massive amount of information that exceeds the capacities of main memory, local disk, and even remote disk (Friedman 2012). According to the National Institute of Standards and Technology (NIST), "*Big Data is a term used to describe the large amount of data in the networked, digitized, sensor-laden, information-driven world*" (Chang and Grady 2015). This definition refers to the bounty of digital data from various data sources in the context of Digital Earth, which focus on big data's geographical aspects of social information, Earth observation (EO), sensor observation service (SOS), cyber infrastructure (CI), social media and business information (Guo 2017; Guo et al. 2017; Yang et al. 2017a, b). Digital Earth data are collected from satellites,

Y. Li · M. Yu · M. Xu · J. Yang · D. Sha · Q. Liu · C. Yang (✉)
Department of Geography and GeoInformation Science, NSF Spatiotemporal Innovation Center, George Mason University, Fairfax, VA, USA
e-mail: cyang3@gmu.edu

H. Guo et al. (eds.), *Manual of Digital Earth*,
https://doi.org/10.1007/978-981-32-9915-3_9

Table 9.1 Definition of the "9Vs" of big data

"V"	Definition
Volume	The vast data size that traditional data storage and computing technologies cannot easily capture, store, manipulate, analyze, manage and present
Variety	The diversity of data formats and sources. The data formats include text, geometries, images, video, sounds or a combination
Velocity	The speed of data production, storage, analysis, and visualization based on advanced development of data collection methods, i.e., the massive number of sensors in the Interest of Things (IoT) and social media networks
Veracity	The varying reliability, accuracy, or quality of data sources
Validity	The accuracy and correctness of Earth data for the intended usage
Variability	The meaning of data continues to change, particularly for Earth data that relies on natural language processing
Vulnerability	Data security is an important part of typical and big Earth data because some geospatial data contain identification information related to people or governments
Volatility	The timeliness and freshness of Earth data
Visualization	Visualization of Earth data is challenging with limited memory, poor scalability and functionality, and various data increasing at a high velocity
Value	Value reflects the tremendous straightforward and potential scientific and social worth based on imaginative insight and analysis results

sensors, simulation models, mobile phones, utilities, vehicles, and social networks in different formats, e.g., imagery, text, video, sound, geometries and combinations of them (Yang et al. 2017a, b). Digital Earth data are naturally big data because of the variety of data sources and enormous data volume.

The increasing availability of big Earth data has provided unprecedented opportunities to understand the Earth in the Digital Earth context. In recent research, big data have been characterized by 5 Vs (volume, variety, velocity, veracity, and value) (Gantz and Reinsel 2011; Zikopoulos and Barbas 2012; Marr 2015). Firican (2017) extended the 5 Vs into big data characteristics including variability, validity, vulnerability, volatility and visualization (as defined in Table 9.1 and further elaborated below).

Volume

The volume of remote sensing imagery collected by satellites and drones easily reaches the TB and PB levels. For example, the Integrated Multi-satellite Retrievals for GPM (IMERG) data product records global precipitation information every half hour, producing up to 3.45 TB data yearly (Huffman et al. 2015). Other location-based data such as social media (e.g., Twitter) and VGI (e.g., OpenStreetMap) are constantly growing.

Variety
Data sources include sensors, digitizers, scanners, numerical models, mobile phones, the Internet, videos, emails, and social networks in the context of Digital Earth. All types of geospatial data require a more effective data structure, framework, index, model, management methodology, and tactics. In addition, these geospatial data are formatted in various data models, e.g., vector and raster, structured and unstructured.

Velocity
The speed of Earth data collection and generation has increased with the development of advanced techniques such as drone observation for disaster monitoring. For example, with the massive number of object-based sensors in the IoT, the data generation of IoT nodes is fast since most sensors continuously generate data in real time.

Veracity
The accuracy of geospatial data varies by data source (Li et al. 2016). Taking precipitation as an example, the quality of remote sensing images such as TRMM and IMERG depends on the sensor configuration, calibration methods, and retrieval algorithms. Precipitation information in MERRA (Modern Era Retrospective-analysis for Research and Applications) data relies on the sophistication of meteorological models. Stationary data collected by rain gauges are more accurate even though they are sparse.

Validity
Similar to veracity, validity concerns the accuracy and correctness of Earth data for the intended usage. In addition to data selection in which data are chosen with appropriate spatial and temporal resolutions and variables for a specific application, data preprocessing, e.g., data augmentation, interpolation, outlier detection, also play an important role in uncovering information from big Earth data. Consistent data quality, common definitions and metadata can benefit the community, resulting in Earth data of high validity.

Variability
Variability refers to the continuous change in the meaning of data in the context of big Earth data, particularly for data that relies on natural language processing. For example, Twitter data emerged as an additional source for natural disaster management (Yu et al. 2018), as tweets posted during disasters can be collected to aid situational awareness. The meaning of words constantly changes over time, for example, the word "Irma" may be a name but started to represent the strongest observed hurricane in the Atlantic in most tweets around October 2017.

Vulnerability
Security is a challenging aspect because some geospatial data contain identifiable information or are sensitive. For example, cellular data have been widely utilized to analyze human activities in smart city applications, however, showing phone numbers may divulge people's private affairs.

Volatility
Volatility refers to the timeliness and freshness of Earth data, i.e., how long the Earth data stay useful and relevant to applications and how long the data should be kept. Due to the velocity and volume of big Earth data, it is impossible to store all the data in a live database without any performance issues. A series of rules should be established for data currency, availability and rapid retrieval (Firican 2017), e.g., historical and less frequently visited Earth data could be archived on a lower-cost tier of storage.

Visualization
Visualization of Earth data is a challenging task with limited memory due to poorly scalable, low-functionality, and high-velocity datasets. Traditional methods may fail to render billions of points, polylines and polygons when visualizing geospatial vector data, therefore graphical methods, e.g., data clustering, parallel coordinates, cone tree or circular network diagrams, should be used to represent Earth data (Firican 2017).

Value
Value presents a low-density pattern in the current big data ecosystem where only a small portion of data is utilized in practice. Earth data occupies 80%+ of our data assets (Dempsey 2012), but most datasets are not excavated and are under-utilized. With appropriate spatiotemporal resolution and analysis methods, the 9Vs have been addressed to obtain actionable knowledge to increase the value of big data.

Data collection strategies, data storage facilities, data analysis methods, and data access services facilitate the transformation from the other 9Vs to the 10th V of value. With the continuing increases in the volume and complexity of data, there are challenges in the life cycle of data management, including data storage, data query, data analysis, data sharing, and many other aspects. Managing big data requires an extensible, interoperable and scalable architecture that supports data storage and analysis. Fortunately, recent years have witnessed the evolution of cloud computing, which brings potential solutions to support the life cycle of big data management.

Cloud computing is a new computing paradigm for delivering computation as a fifth utility, which became popular earlier than big data (Yang et al. 2011a). It has the features of elasticity, pooled resources, on-demand access, self-service and pay-as-you-go characteristics (Mell and Grance 2011) and was termed spatial cloud computing in the context of Digital Earth (Yang et al. 2011a). Big data technologies, e.g., big data storage and big data analytics, evolve and benefit significantly from their integration with cloud computing.

To provide a comprehensive overview of how cloud computing supports big data in the context of Digital Earth, this chapter introduces Digital Earth data sources (Sect. 9.2), data analysis methods (Sect. 9.3), architecture for big data analysis (Sect. 9.4), and cloud computing and its support of big data management (Sect. 9.5). Two examples of EarthCube and Data Cube are introduced in Sect. 9.6 to exemplify cloud-based big data frameworks in the Digital Earth context.

9.2 Big Data Sources

With the advanced developments in Earth observation systems, various Earth data have been gathered at a high velocity from five major sources: (1) remote sensing, (2) in situ sensing, (3) simulation, (4) social media, and (5) infrastructure management (Fig. 9.1). Each covers more than one characteristic of big data. This section discusses the five data sources.

Remote sensing data
Referring to the USGS's definition, remote sensing is the process of detecting and monitoring the physical characteristics of an area by measuring the reflected and

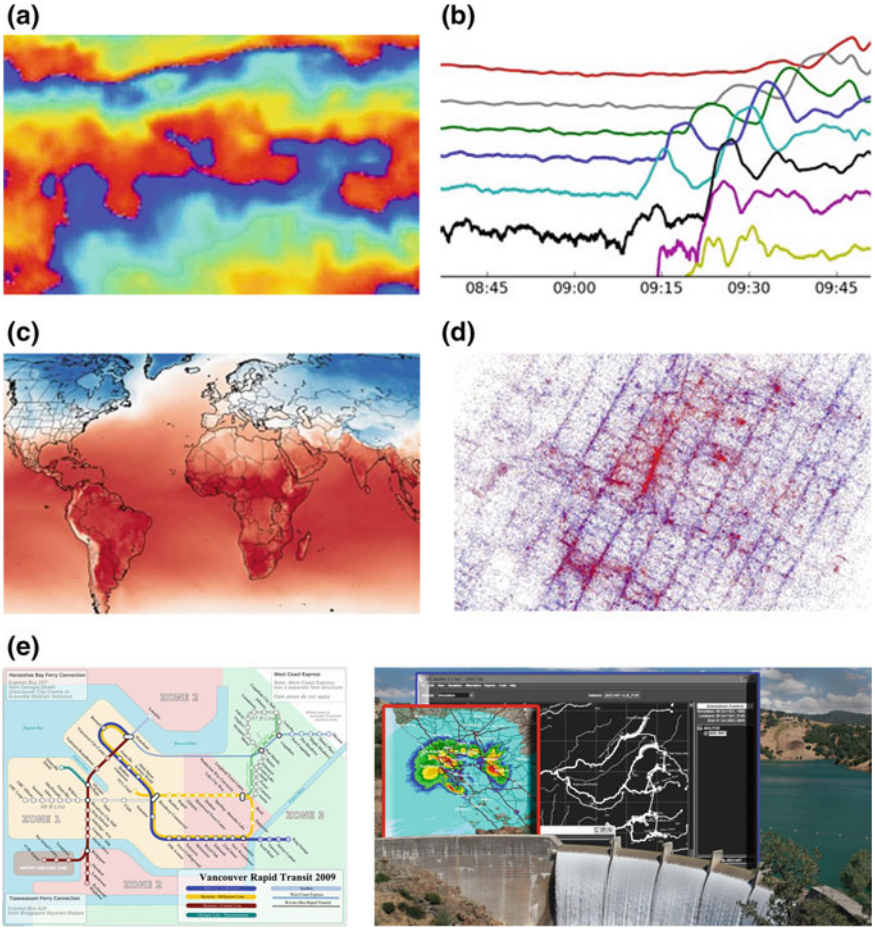

Fig. 9.1 Big Earth data sources: **a** remote sensing data (JPL 2001); **b** in situ data (NOAA 2017); **c** simulation data (Lipponen 2017); **d** social media data (Gundersen 2013); and **e** infrastructure data (Canada Line Vancouver Transit Map 2019; Robert 2000)

emitted radiation at a distance from the targeted area (USGS 2019). Such remotely observed data serve as a vital source for tracking natural phenomena, the growth of a city, changes in farmland or forest, and discovery of the rugged topography of the ocean floor. According to the Earth Observing System Data and Information System (EOSDIS) 2014 statistics, EOSDIS manages over 9 PB of data and adds 6.4 TB of data to its archives every day (NASA 2016). As data precision and density increase over time, data volume increases exponentially. In addition, ongoing international initiatives monitor the Earth in near-real time using satellites to support rapid data collection for quick and effective emergency response (Zhang and Kerle 2008). Remote sensing data are big data due to the big volume, variety, veracity and volatility.

In situ data

According to NOAA (National Oceanic and Atmospheric Administration), in situ data are measurements made at the actual location of the object. In contrast to remote sensing, in situ sensing harvests data directly at the observation location, and often provides continuous data streams to reflect the actual situation with very low latency. Examples of such measurements are (1) tall tower networks (NOAA ESRL/GMD) that provide regionally representative measurements of carbon dioxide (CO_2) and related gases and (2) moored and drifting buoys for marine/ocean data collection. In situ data are big data considering the volume, velocity, and veracity.

Simulation data

Simulation datasets or reanalysis datasets refer to the outputs of Earth models (e.g., climate) based on geophysical principles. By assimilating observations with models, better initial conditions can be leveraged and simulation results can be significantly improved, especially for short-term predictions. Simulation datasets can be used in various applications. For example, the precipitation, evaporation, and runoff from MERRA datasets can drive river flow models and enhance the study of sensitive ecosystems such as estuaries (Rienecker et al. 2011). In addition, the reanalysis winds used in transport models support the evaluation of aerosols. Simulation data are big data due to its volume, variety and validity.

Social media data

In recent years, social media has become one of the most popular sources of big data and provides valuable insights on event trends and people's references. Social networks such as Twitter and Facebook generate a vast amount of geo-tagged data every second and are transforming social sciences (Yang et al. 2017a). Scientists from economics, political science, social science, and geoscience domains utilize big data mining methods to detect social interactions and analyze health records, phone logs, and government records (Balakrishna 2012). For example, in Digital Earth, social media and crowdsourcing data can provide trends of the urban flooding events or wildfire spread, as well as support near-real time situational awareness when other types of data are limited or hard to obtain. However, social media data have high uncertainty and vary in format and quality. Tweet content analysis highly relies on natural language processing, but word meaning constantly changes. Social media

data are big data due to its volume, velocity, variety, veracity, validity, variability and volatility.

Infrastructure data
Infrastructure data serve as a vital data source of Digital Earth information, especially for developing smart cities. For example, basic infrastructure data (e.g., utility, transportation, and energy), healthcare data and governance data (e.g., environmental and construction management) should be proposed, planned and provided by local official departments and business agencies for a smart city (Hashem et al. 2016). Some infrastructure data may contain sensitive information. Taking water distribution management systems as an example, a synthetic data methodology was proposed to reproduce water consumption data according to privacy constraints (Kofinas et al. 2018). With the upgrades in infrastructure, Internet of Things (IoT) data, geo-tagged or geo-referenced data are continuously produced by various devices, sensors, systems and services (Boulos and Al-Shorbaji 2014). In the near future, various applications based on IoT data will benefit individuals and society. For example, near-real time data including temperature and wind information gathered by IoT sensors could support real-time urban microclimate analysis (Rathore et al. 2017). Infrastructure data are big data due to its volume, velocity, variety, veracity, vulnerability, validity and volatility.

Earth data are continuing to grow in volume and complexity. Big data analytical methods are utilized to mine actionable knowledge from big Earth data to convert the 9Vs of Earth data to the 10th V, which is discussed in the next section.

9.3 Big Data Analysis Methods

The advancements in remote sensing, social networking, high-performance simulation modeling and in situ monitoring provide unprecedented big data about our planet. The large volume and variety of data offer an opportunity to better understand the Earth by extracting pieces of knowledge from these data. This section discusses data analysis methods from the three aspects of data preprocessing, statistical analysis and nonstatistical analysis. The characteristics, applications, and challenges of these methods are introduced below.

9.3.1 Data Preprocessing

Real-world data are usually incomplete, noisy and inconsistent due to data collection limitations and sensor issues. Raw data may contain errors or outliers, lack specific attributes or have discrepancies in the descriptions. Therefore, data preprocessing (e.g., data cleaning, fusion, transformation, and reduction) are required to remove noise, correct data, or reduce data size.

Low-quality values (missing values, outliers, noises, inconsistent values) in raw data are often removed or replaced with user-generated values, e.g., interpolation values. The missing value is usually represented with a symbol (e.g., N/A) in raw data and easily recognize. Outliers and inconsistent values are hidden in the raw data and can be detected through statistical analysis. Taking water usage behavior analysis as an example, data preprocessing is necessary to turn smart water meter data into useful water consumption patterns because IoT sensors may fail to record data (Söderberg and Dahlström 2017).

Data transformation also plays an essential role in data preprocessing. Multiple Digital Earth data, e.g., climate data, soil moisture data, crop data, are converted to monthly z-score data before analysis to eliminate the seasonal trends that usually make the patterns of interest undiscoverable. Aggregation, another important data transformation method, groups data based on numerical attributes (Heuvelink and Pebesma 1999). In the Earth science domain, aggregating raw data to the county or state levels could uncover essential patterns for decision making, urban planning, and regional development.

Another trend in Digital Earth data analysis is multisource data fusion, which provides comprehensive data retrieved from several data sources. Generally, vector and raster data store Earth information with different spatial-temporal resolutions; thus, data must be converted to the same resolution by interpolating the lower resolution data or aggregating the higher resolution data for intelligent analysis to investigate scientific questions at a specific scale. Sharifzadeh and Shahabi (2004) introduced a spatial aggregation method that takes the sensor data distribution into account. Spatial interpolation is interpolation of point and areal data. Point interpolation is applied to contour mapping and areal interpolation is used in isopleth mapping (Lam 1983). In addition to spatial interpolation, temporal interpolation predicts values between timestamps (Lepot et al. 2017).

9.3.2 Statistical Analysis

In the era of big data, statistical analysis is a common mathematical method of information extraction and discovery. Statistical methods are mathematical formulas, models, and techniques used to find patterns and rules from raw data (Schabenberger and Gotway 2017). Data mining is the process of discovering patterns from large datasets involving statistical analysis. Through data mining, historical data can be transformed into knowledge to predict relevant phenomena. Both traditional statistics and data mining methods are discussed in this section. These methods include but are not limited to regression analysis, spatiotemporal analysis, association rules, classification, clustering, and deep learning.

Regression analysis

Regression models the relationships between a dependent variable and one or more explanatory variables (Yoo et al. 2014; Anderson 2015) by estimating the values of

a dependent variable when the values of the independent variables are known and the relationships exist. Regression models describe the strength or weakness of the relationship between several variables. For example, Blachowski (2016) proposed a weighted spatial regression method that identified four significant factors inducing land subsidence: thickness, inclination, the depth of coal panels, and the slope of the surface. There are challenges in spatial data analysis using regression methods, especially for situations that are complicated enough to result in serious residuals in the regression models.

Spatiotemporal analysis
Spatiotemporal data analysis investigates the trajectories and trends of spatiotemporal data. Li et al. (2016) investigated the spatiotemporal trends in the fluctuations of housing price data. Spatial data analytics and modeling techniques were used to identify the spatial distribution of housing prices at the micro level and explore the space-time dynamics of residential properties in the market, as well as the detected geographic disparity in terms of housing prices. Rahman and Lateh (2017) analyzed the temperature and rainfall time series data from 34 meteorological stations distributed throughout Bangladesh over 40 years (1971–2010) to statistically evaluate the magnitude of temperature and rainfall changes across space and time. Spatiotemporal analysis is still in its initial stage of development. Challenging questions remain, such as what kinds of patterns can be extracted from time series data and which methods and algorithms should be applied.

Association rule
Association rule learning is the process of discovering strong relationships between variables, i.e., rules, in a large database using measurements of support and confidence (Agrawal et al. 1993). In Digital Earth, Yang (2011b, 2016) applied association rules to mine the variables of Atlantic hurricanes from 1980 to 2003 and discovered a combination of factors related to rapid intensification probability, the low vertical shear of the horizontal wind (SHRD = L), high humidity in the 850–700 hPa range (RHLO = H), and tropical cyclones in an intensification phase (PD12 = H). Compared with traditional statistical methods, the rule-based mining method can find combinations of factors instead of a single factor related to an event.

Classification
Classification learning is the task of mapping input variables to discrete output variables called labels or categories, for example, 'building' or 'road.' It is the process of recognizing, differentiating and understanding objects. Support Vector Machine (SVM) is a classical classification algorithm in which a kernel-based metric is used to differentiate objects. Jiang et al. (2018b) integrated the ranking support vector machine (RankSVM) model from the computer science community with ocean data attributes to support data ranking in ocean portals. An SVM model is also used to predict geological lithofacies from wireline logs.

Clustering
Clustering is the process of splitting a set of objects into closely related groups, and each group is regarded as a cluster. Objects falling in the same cluster are more

similar to each other than those in other clusters. In Digital Earth, clustering plays an important role in pattern analysis. Hong and O'Sullivan (2012) clustered empirical datasets in Auckland, New Zealand for ethnic residential cluster detection, which is useful to understand contemporary immigrants and ethnic minorities in urban areas. Zhao et al. (2017) proposed a method to generate urban road intersection models from low-frequency GPS trajectory data. These patterns identified from empirical data are crucial for urban transportation planning and management.

Deep learning
As a new paradigm of machine learning, deep learning has achieved remarkable success in discovery of implicit knowledge (LeCun et al. 2015). In Digital Earth, deep learning algorithms have been adopted to solve domain problems. For example, Guo and Feng (2018) used multiscale and hierarchical deep convolutional features to assign meaningful semantic labels to the points in a three-dimensional (3D) point cloud, which is essential for generating 3D models. Li and Hsu (2018) proposed a deep learning approach to automatically identify terrain features (i.e., sand dunes, craters) from remote sensing imagery. Compared with traditional induction-based approaches, the deep learning approach could detect diverse and complex terrain features more accurately and process massive available geospatial data more efficiently.

9.3.3 Nonstatistical Analysis

In addition to statistical analysis, nonstatistical analysis methods also play an essential role in helping us descriptively understand Earth phenomena. This section introduces two representative models in Digital Earth, linked data and 3D city modeling.

Linked data
Linked data are structured data in which datasets are interlinked in the collection, which is useful for semantic queries and reasoning (Bizer et al. 2011). With linked data, data are sharable, and the relationships among the data are recorded. Standard web technologies such as RDF (Resource Description Framework) provide a way to build shareable linked data. In Digital Earth, heterogeneous Earth data (multidisciplinary, multitemporal, multiresolution, and multilingual) can be integrated based on linked data principles for decision making and knowledge discovery (Vilches-Blázquez et al. 2014). For example, Mc Cutchan (2017) proposed a structure of embedding geographic data into linked data and forecasted spatial phenomena with associated rules extracted from the linked data.

3D city modeling
A trend in Digital Earth analysis is to build a real 3D model with the aid of a computer, especially for cities where most human activities occur. 3D models provide real three-dimensional information for analysis, going beyond simple visualization of 3D objects. 3D informatics has become a cornerstone for a series of city-related

applications such as urban planning, skyline analysis, crisis and disaster management, route selection and navigation (El-Mekawy 2010). For example, Amirebrahimi et al. (2016) assessed and visualized flood damage using 3D urban modeling and a building information model (BIM), improving the resilience of the community to floods using detailed 3D information.

Knowledge distillation from Earth data has demonstrated excellent improvements in our understanding of the planet we live. As Earth data increase faster than ever, state-of-the-art analysis methods should be developed to handle the increasingly complicated spatiotemporal data. In addition, an extensible, interoperable and scalable architecture is a prerequisite for massive geographic data analysis, and we present a big data analysis architecture in the next section.

9.4 Architecture for Big Data Analysis

To support Earth data access/query/analysis in a reasonable response time, it is crucial to build a sophisticated analytical platform with robust architecture to reveal insights from the data (Yang et al. 2017a, b). Generally, the architecture of analytical platforms consists of a data storage layer, a data query layer, a data processing layer, and a visualization layer (Fig. 9.2).

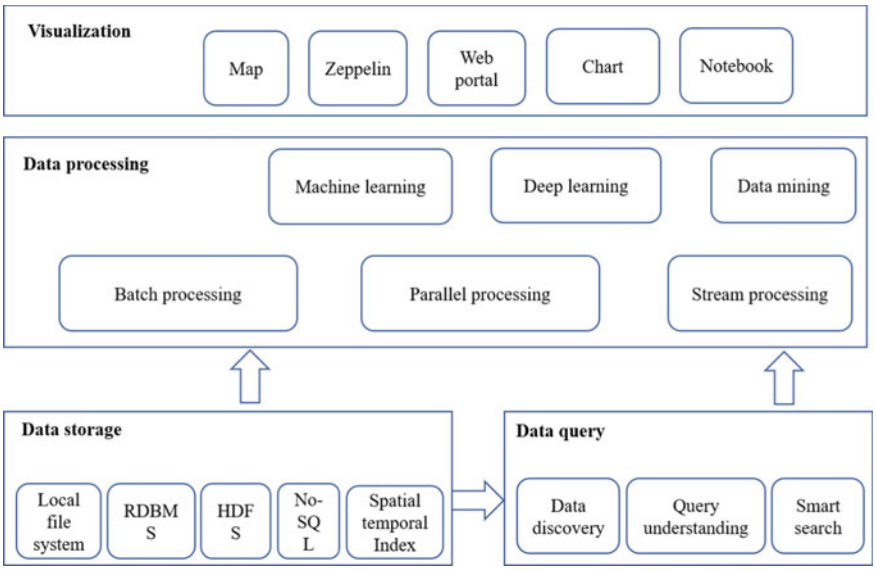

Fig. 9.2 Architecture for big data analyses

9.4.1 Data Storage Layer

Digital Earth is heavily involved in processing big data from Earth observations and model simulations, and these vast amounts of high-dimensional array-based spatiotemporal data pose challenges for data storage (Li et al. 2017b). Customizations are indispensable in integrating advanced data storage technologies with big Earth data storage, as general distributed file systems such as Hadoop are not designed to store spatiotemporal data.

A robust and stable data storage framework is the foundation of the data analysis architecture. A series of research efforts focused on optimizing spatiotemporal data storage in distributed file systems or databases. For example, Hu et al. (2018a) reorganized NetCDF (Rew and Davis 1990), a standard data format for array-based raster data, into CSV files and deployed them within SciDB (Cudre-Mauroux et al. 2009), a scalable multidimensional array clustering database. Zhao et al. (2010) converted NetCDF data into CDL (network Common data form Description Language) files and distributed them on HDFS (Hadoop Distributed File System). MERRA data, which store reanalysis Earth climatic variables in NetCDF format, were transformed into Hadoop Sequence Files to be processed by standard MapReduce functions (Duffy et al. 2012). Li et al. (2015) decomposed array-based raster data and stored them with HBase, a NoSQL database built upon HDFS in a cloud computing environment for efficient data access and query.

To enable efficient big data query, logical query capabilities have been proposed to support spatiotemporal query of array-based models such as SciHadoop (Buck et al. 2011). A spatiotemporal index was designed to efficiently retrieve and process big array-based raster data using MapReduce and a grid partition algorithm atop the index to optimize the MapReduce performance (Li et al. 2017a). SciHive was developed as an extension of Hadoop Hive, mapping arrays in NetCDF files to a table and calculating the value range for each HDFS to build a distributed adaptive index (Geng et al. 2013, 2014). Malik (2014) introduced a Z-region index into GeoBase to facilitate array-based data storage.

9.4.2 Data Query Layer

To help data consumers efficiently discover data from the massive available Earth data, the Digital Earth communities have built various data portals to improve the discovery, access, and usability of Earth data. The portals are normally supported by text search and spatiotemporal search and include the GeoPortal,[1] GeoNetwork[2] Spatial Web Portal (Xu et al. 2011), Global Earth Observation System of Systems GEOSS Clearinghouse (Liu et al. 2011; Nativi et al. 2015; Giuliani et al. 2017),

[1] https://www.geoportal.gov.pl/.

[2] https://geonetwork-opensource.org/.

GeoSearch (Lui et al. 2013) and many others. For example, GEOSS is a cloud-based framework for global and multidisciplinary Earth observation data sharing, discovery, and access (Nativi et al. 2015). In the framework, datasets or workflows are registered into shared collections or global catalogs, allowing for end users to search for workflows and datasets across multiple granularity levels and disciplines.

In addition, open-source information retrieval frameworks, e.g., Apache Lucene or its variants such as Solr and Elasticsearch (McCandless et al. 2010), were adopted to establish an Earth data portal instead of implementing a search engine for Earth data from scratch. Lucene uses the Boolean model and the practical scoring function to match documents to a query. Solr and Elasticsearch improve the Lucene index to enable big data search capabilities. The Physical Oceanography Distributed Active Archive Center (PO. DAAC) serves the oceanographic community with 514 collection level datasets and massive granule level data atop Solr (Jiang et al. 2018a). Elasticsearch is the fundamental component of NOAA's OneStop portal in which data providers manage data and metadata with increased discoverability and accessibility.

However, solely relying on open source solutions is insufficient for Earth data discovery because these solutions only rely on a keyword-based relevance score for ranking and ignore other user preferences, e.g., data processing level, sensor type. A few related research efforts have been conducted in the Earth science domain to make data search engines smarter and more intelligent. For example, an algorithm combing Latent Semantic Analysis (LSA) and a two-tier ranking was reported to build a semantic-enabled data search engine (Li et al. 2014a, b). Jiang et al. (2018a) developed a smart web-based data discovery engine that mines and utilizes data relevancy from metadata and user behavior data. The engine enables machine-learned ranking based on several features that can reflect users' search preferences.

9.4.3 Data Processing Layer

Data processing layer is a core component of the data analytics architecture. To analyze terabyte and petabyte datasets with low time latency, even in a real-time manner, sophisticated parallel computing algorithms and scalable computing resources are required in the big data processing framework (Yang et al. 2015a). Advanced open-source parallel computing solutions, e.g., Hadoop MapReduce, Spark, and their variants in the Earth data domain have been leveraged to support data analysis and mining tasks with better performance.

Hadoop MapReduce is a high-performance batch processing parallel framework that solves large computational problems on distributed storage systems (White 2012). It transfers the algorithm code to data nodes rather than moving data blocks to a compute node to avoid I/O bottlenecks. Spark enables high-performance data analysis with in-memory computing. An in-memory data structure called the Resilient Distributed Dataset (RDD) manages datasets distributed in a Spark cluster (Zaharia et al. 2012). However, the original distributed frameworks have limitations on big spatiotemporal data processing. Yu et al. (2015) noted that the system scalability

for spatiotemporal data and interactive performance are the two main challenges for big Earth data processing. To solve these problems, scientists and engineers have customized open-source solutions for spatiotemporal data analysis.

SpatialHadoop is a MapReduce-based framework with native support for spatial data including a simple spatial high-level language, a two-level spatial index structure, a fundamental spatial component built on the MapReduce layer and three basic spatial operations (range query, k-NN query, and spatial link) (Eldawy and Mokbel 2015). GeoSpark provides operational tools for spatial big data processing based on Spark (Yu et al. 2015). The Spatial Resilient Distributed Datasets (SRDDs) structure represents spatial data blocks in memory and index objects using quad-tree and r-tree in each RDD partition (Lenka et al. 2016). ClimateSpark integrates Spark SQL and Apache Zeppelin to develop a web portal that facilitates interaction among climatologists, climate data, analytical operations and computing resources (Hu et al. 2018b). As an extension of Scala, GeoTrellis supports high-performance raster data analysis. GeoMesa provides spatiotemporal data persistence on Hadoop and column-family databases (e.g., Accumulo, HBase), as well as a suite of geospatial analytical tools for massive vector and raster data (Hughes et al. 2015).

As described in this section, a service-oriented, scalable architecture usually contains three major layers to provide desirable functionalities and capabilities: (1) the bottom data storage layer provides physical data storage, (2) the data query layer enables data discovery capabilities with proper functionality and interoperability, and (3) the data processing layer supports extensible, interoperable and scalable analytical functionalities based on open source solutions and their variants from the geoscience communities. With the architecture, big Earth data could be accessed and analyzed with low time latency or even in real time. However, it is challenging to set up such architecture and share data stored inside them due to the requirements of storage resources, computing resources, complicated configurations, and domain knowledge. Fortunately, the paradigm of cloud computing, discussed in the next section, brings potential solutions to ease the process of analytical framework setup and data sharing.

9.5 Cloud Computing for Big Data

9.5.1 Cloud Computing and Other Related Computing Paradigms

Grid computing and High Performance Computing (HPC) have been utilized for big data analytics. Grid computing, a distributed system of computer resources, performs large tasks using loosely coupled computers in a distributed system (Hamscher et al. 2000). The European Data Grid project utilizes grid computing to support exploration of multi-petabyte datasets (Segal et al. 2000) and the TeraGrid GIScience gateway utilized grid computing to perform computationally intensive geographical analytics

(Wang and Liu 2009). HPC uses supercomputers to run applications in parallel, efficiently and quickly, and is used in the PRACE project to serve European scientists with high-performance computing capabilities to conduct research (Hacker et al. 2010).

Cloud computing virtualizes computer resources as a resource pool to provide computing resources over the network by optimizing resource usage in terms of the CPU, RAM, network, and storage. Cloud computing has intrinsic connection to the Grid Computing paradigm (Foster et al. 2008) in that both are distributed computing systems. Cloud computing relies on remote servers whereas grid computing connects servers or personal computers over a common network using a Wide Area Network (WAN) to perform parallel tasks (Foster et al. 2008). Compared with HPC, cloud computing is cost effective and easy to use. Although cloud computing can provide high performance computing capability, HPC is irreplaceable for some applications since supercomputers are required to process very complicated processes such as climate simulations. In addition, resources in cloud computing are controlled by the service providers and users have limited controls.

In addition to cloud computing, other new computing paradigms have emerged to build a comprehensive and economic computing framework. For example, edge computing can process data at the edge of network due to the advancement of the IoTs. IoT applications usually produce a massive amount of streaming data and require near-real time response; thus, cloud computing alone is not an optimal solution for data collection and analysis for such real-time applications. In edge computing, edge nodes serve as data providers and consumers to protect data privacy and make full use of the computing capacity of edge nodes. Less data is transferred to the cloud computing platform after data preprocessing in edge nodes, reducing the response time and bandwidth cost (Shi et al. 2016).

Mobile computing with portable computing nodes has become an important computing paradigm with the improvements in the computing and storage capacity of smart devices such as smartphones and tablets (Qi and Gani 2012). Although services provided by mobile computing are not as reliable as edge computing and cloud computing due to the restrictions in battery volume and network connection, mobile computing can collect data and reach end users where cloud computing and edge computing are inaccessible.

These computing paradigms have advantages and disadvantages, and can be integrated to complement each other and provide reliable and effective data storage and processing frameworks according to the data characteristics and computing requirements. Cloud computing and big data are the two most important technologies in Digital Earth. The following section discusses the utilization of cloud computing to support big data management in Digital Earth.

9.5.2 *Introduction to Cloud Computing*

As a new computing paradigm, cloud computing delivers scalable, on-demand, pay-as-you-go access to a pool of computing resources (Mell and Grance 2011; Yang et al. 2011a). Practically, cloud computing aims to maximize the utilization rate of physical resources and provide virtual resources to aid applications and services. Cloud computing relies on several technologies including virtualization, network security, and high availability to provide services over the network. These technologies make it easier, more efficient, and more economical to set up architecture for big data analysis.

Virtualization is the fundamental technology for cloud computing, which abstracts an application, operating system, or data store from the underlying hardware or software. Virtualization creates a "layer of abstraction" between the physical systems and a virtual environment in the virtualization process (Big Data Virtualization). Server virtualization optimizes the use of redundant computing and storage resources by virtualizing distributed computer resources (e.g., CPU, RAM, Network, and Disk) and managing them in the same resource pool. With virtualization, cloud computing can provide on-demand big data services and support big data technologies including big data storage, process, analysis, visualization, and remote collaboration (Fig. 9.3). Virtualizing big data resources as a pool serves as a user-friendly interface and makes big data analytics accessible to end users.

As one of the cloud solutions, public clouds are the most accessible cloud computing services offered by third-party providers (e.g., Amazon Web Services (AWS), Microsoft Azure, Alibaba Cloud) over the Internet. Public clouds are available to the public and may be offered on a pay-per-usage model (Li et al. 2010). In contrast to public clouds, private clouds are dedicated for use inside an organization. Private cloud resources can be managed by an organization or by a third-party vendor,

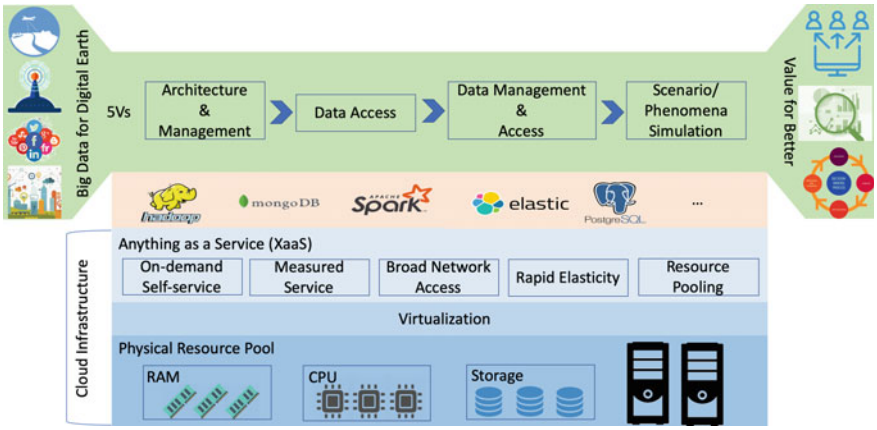

Fig. 9.3 Cloud computing for big data analysis

regardless of the physical location of the resources (Dillon et al. 2010). The computing resources in a private cloud are isolated and delivered via a secure private network.

The advanced features of auto scaling and load balancing through resource monitoring further maximize the capability of cloud computing resources. Based on the individual performance of a machine, autoscaling can be applied to allow for some servers to rest during times of low load to save electricity costs and automatically add more instances during times of high demand. In addition, load balancing improves the distribution of workloads across multiple instances to optimize resource use, minimize response time, and avoid overloading any single instance.

9.5.3 Cloud Computing to Support Big Data

Cloud computing combines distributed computing resources into one virtual environment, providing big data analytics and solutions during the life cycles of big data. Three main categories of cloud computing services are (1) Infrastructure as a Service (IaaS), (2) Software as a Service (SaaS), and (3) Platform as a Service (PaaS). Together with Data as a Service (DaaS), Model as a Service (MaaS; Li et al. 2014a, b) and workflow as a service (WaaS; Krämer and Senner 2015), cloud computing offers big data researchers the opportunity of anything as a service (XaaS; Yang et al. 2017b).

Cloud Storage for Big Data Storage
The characteristics of big data in high volume lead to challenges for data storage. Cloud computing's potential for unlimited storage support helps solve the volume challenge of big data, as the cloud provides virtually customizable storage with elastically expandable and reducible size. An alternative solution is Data Storage as a Service (DSaaS) enabled by block storage, which is the capability of adding external storages as "blocks". With block storage, it is possible to enlarge the storage size without physically loading hard drives. Virtually unlimited scalable storage offered by cloud computing grants users the capability of dynamic adjustment to satisfy the storage requirements of data with high volume and velocity. The modularized virtual resource offers effortless data sharing within production environments by allowing for an external data block to be detached and remounted from one machine to another. External data storage can be automatically backed up to prevent users from losing data, and backups that are securely saved at the back-end server can be easily transferred and restored. In addition, information security is guaranteed because the physical location cannot be obtained from the disk drive (Mayama et al. 2011).

Cloud Computing for Big Data Processing
Processing large volumes of data requires dedicated computing resources, e.g., faster CPUs and networks and larger disks and RAMs (Yang et al. 2017b). Cloud computing

provides on-demand resources and delivers configurable resources including mountable external storage spaces, computing resources (CPU, RAM), and network services. Traditionally, a computer uses approximately two-thirds of the power of a busy computer (JoSEP et al. 2010), and cloud computing has the potential to provide on-demand computing resources. Isolated virtual structures have been created for big data systems to enhance system stabilities, which can be easily managed in different file systems and replicated through backup images to provide fast configuration recovery. The ability to replicate environments automates the expansion of compute nodes in virtual machine clusters, thereby efficiently utilizing resource pools to support big data analytics. With the foundational support of storage for big data, data processing inherits the advantages of fast data acquisition and relocation.

Although cloud computing could serve as an excellent infrastructure option for big data processing, several aspects should be considered to minimize the bottleneck effect for the general processing speed, such as the choice of cloud volume type according to I/O demand and cloud bandwidth selection according to application requirements.

Cloud Computing for Big Data Analytics

Popular big data analytical platforms such as Apache Hadoop are traditionally installed on physical machine clusters, resulting in a waste of computing resources due to hardware redundancy (CPU and RAM). With the virtual clusters provided by cloud computing through virtualization technology, distributed analytical platforms can be migrated to the virtual clusters from physical machine clusters, optimizing the usage of computing resources in an efficient manner.

With the aid of autoscaling and load balancing, deploying on-demand and scalable big data analytical platforms could easily provide resilient analytical frameworks and minimize waste of computing resources. Autoscaling supports parallel algorithms on distributed systems and architectures for scalability. It allows for the expanded resources to function when the algorithms or programs are enabled with parallel computing capability. Without it, public cloud providers such as AWS could not offer automatic scalability (JoSEP et al. 2010). The load balancer distributes workloads among virtual clusters and triggers autoscaling functions when analytics require higher computing configurations. The virtual system as a whole could dynamically fit higher computing requirements by launching more virtual duplications as needed. The load balancer acts as a virtual network traffic distributor and can be optimized to better allocate overall resources.

Cloud Computing for Big Data Sharing and Remote Collaboration

Traditional deployment of big data systems requires complicated settings and efforts to share data assets. It lacks access control and often leads to data security and data privacy issues. Cloud computing enhances the sharing of information by applying modern analytical tools and managing controlled access and security (Radke and Tseng 2015). Virtualization enables different parties to share data assets to achieve various goals and objectives under a centralized management system. With the support of cloud computing, it is possible to flexibly share data and remotely collaborate, which involve interdisciplinary collaborations and advanced workflows. Though data

sharing, computational resource sharing, and production environment sharing, cloud computing can potentially be used to build a perceptual environment to support various businesses and applications (Li et al. 2015). Unfortunately, workflow sharing remains challenging due to domain boundaries (Yang et al. 2017b).

9.6 Case Study: EarthCube/DataCube

Big data and cloud computing enable Earth scientists and application developers to create web-accessible frameworks and platforms to efficiently store, retrieve and analyze big Earth data. In the Earth science domain, scientists have proposed a series of data models, frameworks, and initiatives to ensure the success of heterogeneous data sharing and analysis. For example, a 10-year framework initiative on sustainable consumption and production from 2013 to 2023 was launched by the United Nations Environmental Program (UNEP). The Future Earth framework, an international research program in the environmental science community, serves as an evolving platform to support transitions toward sustainability (Lahsen 2016). Microsoft's Eye-On-Earth platform aids climate change research in several European countries by collecting and sharing water and air quality data (Microsoft 2011). As part of the European program to monitor the Earth, the Copernicus Data and Information Access Service (DIAS) platform collects and processes data from remote and in situ sensors and provides reliable information covering six thematic areas including land, ocean, atmosphere, climate, emergency, and security (Bereta et al. 2019). Through its Exploitation Platforms (EP) initiative, the European Space Agency (ESA) built several cloud-based Thematic Exploitation Platforms (TEPs) in a preoperational phase for geo-hazard monitoring and prevention (Esch et al. 2017). The CASEarth Poles comprise a comprehensive big data platform of the three poles including the Arctic, Antarctic and the Tibetan plateau within the framework of the "Big Earth Data Science and Engineering" program of the Chinese Academy of Science (Guo et al. 2017). One of the current initiatives is the NSF EarthCube originated from the NSF GEO Vision report (NSF Advisory Committee for Geosciences 2009). In this section, we introduce the EarthCube project and a big data infrastructure, Data Cube, as two cases of big data and cloud computing in the context of Digital Earth.

9.6.1 EarthCube

NSF EarthCube involves (1) Building Blocks (BBs), to develop novel infrastructure capabilities and demonstrate their value in a science context; (2) Research Coordination Networks (RCNs), to engage the science community around joint goals; (3) Conceptual Designs (CDs), to develop broad architecture design and explore integrative systems; (4) Integrative Activities (IAs), to explore concepts for the design of an enterprise architecture, and (5) Data Infrastructures (DIs) to lay the groundwork

for shared data. The EarthCube concept originated from the NSF GEO Vision report (NSF Advisory Committee for Geosciences 2009), which was issued by the Advisory Committee for NSF's Geosciences Directorate (GEO) and identified the future focus of the Earth science community as '*fostering a sustainable future through a better understanding of our complex and changing planet*.' To achieve the GEO vision, the GEO and Office of Cyberinfrastructure (OCI) jointly launched the EarthCube (NSF 2011) initiative as a driving engine to build a geospatial cyberinfrastructure (similar to a Digital Earth infrastructure, Yang et al. 2010) to (1) understand and forecast the behavior of a complex and evolving Earth system; (2) reduce vulnerability and sustain life; and (3) train the workforce of the future.

EarthCube (2012) is targeted at (1) transforming the conduct of data-enabled geoscience-related research, (2) creating effective community-driven cyberinfrastructure, (3) allowing for interoperable resource discovery and knowledge management, and (4) achieving interoperability and data integration across disciplines (Fig. 9.4).

In addition, EarthCube is evolving within a rapidly growing, diverse, and wide-ranging global environment. In addition to the collaboration within EarthCube, there are other contributing entities ranging from individual data sets and software applications to national and international cyberinfrastructure systems. The NSF has funded the development of EarthCube through individual EarthCube awards since 2013. In 2016, the NSF awarded 11 new EarthCube activities, for a total of 51 awards. A sampling of efforts in EarthCube that benefit from big data and cloud computing are introduced below.

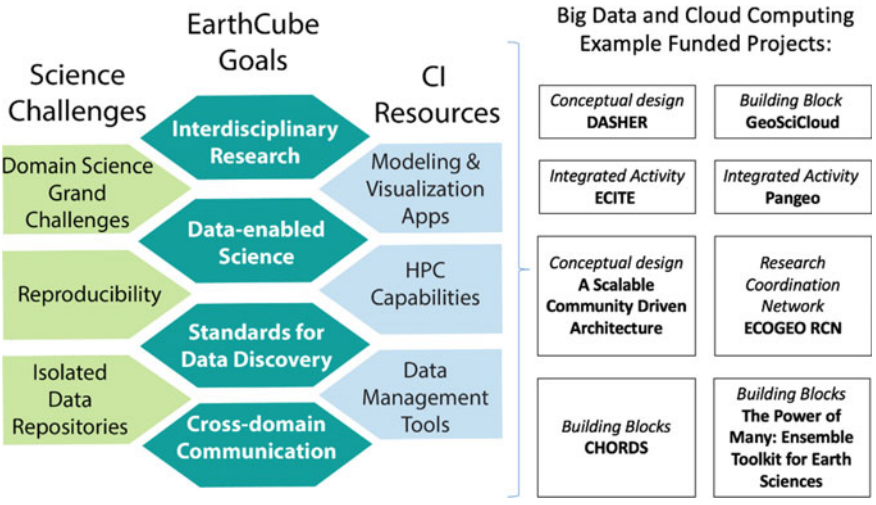

Fig. 9.4 Examples of related projects (derived from EarthCube goals, EarthCube Office 2016)

Conceptual design DASHER
Yang et al. (2015b) proposed a conceptual EarthCube Architecture, DASHER (Developing a Data-Oriented Human-Centric Enterprise Architecture for EarthCube), to support EarthCube and facilitate data communication and social collaboration in pursuit of collaborative Earth sciences research. The final product is a four-volume report containing different viewpoints that describe EarthCube architecture from different conceptual perspectives such as capabilities, operations, services, and projects. It provides a comprehensive conceptual reference for developing a detailed and practical architecture to address the requirements of the EarthCube community. DASHER was one of the first projects funded by EarthCube to design the conceptual framework integrating computational resources and big data sources.

Building Block GeoSciCloud
GeoSciCloud (Deploying Multi-Facility Cyberinfrastructure in Commercial and Private Cloud-based Systems) investigated two medium-size NSF funded data centers to deploy data collections with cloud-based services in different environments to assess feasibility and impact (EarthCube 2019). These environments include (1) commercial cloud environments offered by Amazon, Google, and Microsoft and (2) NSF-supported extensive computing facilities that are just beginning to offer services with characteristics of cloud computing.

GeoSciCloud helps EarthCube compare and contrast these three environments (the Extreme Science and Engineering Discovery Environment (XSEDE), commercial cloud, and current infrastructure) in the massive data ingestion to the cloud, data processing time, elasticity, the speed of data egress from multiple environments, overall costs of operation, interoperability, and reliability of real-time data streaming.

Integrated Activity ECITE
The EarthCube Integration and Test Environment (ECITE) is an outgrowth of activities of the EarthCube Testbed Working Group. The ECITE approach focuses on integrating existing effective technologies and resources as well as capabilities built by the EarthCube community using a cloud platform to provide a federated and interoperable test environment (EarthCube 2016). ECITE engages scientists and technologists from multiple disciplines and geographic regions across the Earth science community to develop requirements, prototype, design, build, and test an integration test-bed that will support cross-disciplinary research. The hybrid federated system will provide a robust set of distributed resources including both public and private cloud capabilities. This research addresses timely issues of integration, testing and evaluation methodologies and best practices with a strong interoperability theme to advance disciplinary research through the integration of diverse and heterogeneous data, algorithms, systems, and sciences.

Integrated Activity Pangeo

Pangeo[3] (an open-source big data climate science platform) integrates a suite of open-source software tools that can tackle petabyte-scale Atmosphere/Ocean/Land/Climate (AOC) datasets. Pangeo aims to cultivate an ecosystem in which the next generation of open-source analysis tools for ocean, atmosphere and climate science can be developed, distributed, and sustained. These tools must be scalable to meet the current and future challenges of big data, and the solutions should leverage the existing expertise outside of the AOC community. The resulting software improvements contribute to upstream open source projects, ensuring the long-term sustainability of the platform. The result is a robust new software toolkit for climate science and beyond. This toolkit will enhance the Data Science aspect of EarthCube. Implementation of these tools on the cloud was tested, taking advantage of an agreement between commercial cloud service providers and the NSF for big data solicitation.

9.6.2 Data Cube

The term 'data cube' was originally used in Online Analytical Processing (OLAP) of business and statistical data but has more recently been used in Earth domains as an approach to manage and analyze large and rapidly growing datasets. In Digital Earth, a data cube represents a multidimensional (n-D) array that stores gridded data or array-based data produced by remote sensing and simulation (Zhang et al. 2005). A data cube can be based on regular or irregular gridded, spatial and/or temporal data with multiple parameters. To support the management, sharing, and serving of Digital Earth data, tools and models, different cyberinfrastructures have been developed based on data cubes. Examples include the EarthServer that provides data cube services for Earth observations based on the RASDAMAN array database (Baumann et al. 2016). Another example is the Earth Observation Data and Processing Platform developed by the European Commission to integrate and analyze the combination of satellite and in situ Earth observations for sustainable development goals (Soille et al. 2016). The Committee on Earth Observation Satellites (CEOS) provides a data processing infrastructure based on data cubes to support Earth science objectives in developing countries, with a focus on remote sensing data (Nativi et al. 2017). The platform automatically ingests different remote sensing data into an N-dimensional data array.

Challenging issues in providing data services in data cube infrastructure include interoperability, rapid data access and transfer, and real-time processing and analysis (Strobl et al. 2017). Interoperability issues occur because datasets from various sources can have distinct parameterizations, spectral band definitions, projections, file formats, and database structures. One solution is to standardize the preprocessing procedure before storage and sharing with the community. The Open Geospatial

[3] http://pangeo.io/.

Consortium (OGC) Sensor Web Enablement (SWE) Common Data Model (CDM) defines important element parameterization (Robin 2011). Datasets must be represented along different predefined dimensions of the data cube, including space, time, and parameter properties. For projection difference issues, OGC recently developed the Discrete Global Grid System (DGGS) to optimize the loss of geospatial data during the reprojection process and seamlessly integrate GIS data from various sources (Stefanakis 2016). Data cube infrastructures also require rapid data access and transfer and real-time processing and analysis. Functionalities for user interactions must be built for various user demands, including file manipulation, data preprocessing, and analysis. These functionalities should also meet the standards of geographical information processing in OGC Web Coverage Services (WCS), geographic information analysis in the OGC Web Coverage Processing Service (WCPS), and format-independent data cube exchange in the OGC Coverage Implementation Schema (CIS).

Cloud computing and big data frameworks could enhance the data cube archive, visualization, and analysis in many ways to meet the needs of big Earth data knowledge mining. In cloud computing, storage is a virtual resource that can be attached and scaled on demand. By leveraging cloud computing and big data frameworks, visualizing data cubes and performing complicated spatiotemporal queries are more effortless than ever before (Zhizhin et al. 2011). One example of data cube visualization in the Earth science domain is the EOD[4] (Earth Observation Data Cube), which enables advanced data access and retrieval capabilities for the European coverage of Landsat-8 and the global coverage of Sentinel2 data. It aims to improve the accessibility of Big Earth data and offers more than 100 TB of Atmosphere, Land and Ocean EO products, demonstrating satellite data in the context of a virtual globe. The ESDC[5] (Earth System Data Cube) is another example of climate data cube visualization and analysis that aims to develop an environment to tap into the full potential of the ESA's Earth observations and integrate with the Biosphere-Atmosphere Virtual Laboratory (BAVL) analysis environment. The use of cloud computing technologies in big data visualization enables a massive amount of end users to explore data online at the same time with very low latency.

Data cube partition and parallel query could be achieved by utilizing distributed big data frameworks, which are faster and easier than traditional noncluster methods. Pagani et al. combined the data cube concept with cloud computing to manage and analyze large Earth datasets and observed better outcomes than traditional file-based approach (2018). Open Data Cube[6] is another example of the utilization of advances in cloud computing, providing free and open technologies to end users without local infrastructure. Thus, developing countries can access data and computing resources to build applications that aid decision making. The Australian Geoscience Data Cube (AGDA) solves similar problems of data sharing. It makes more than three decades of satellite imagery available for the first time, spanning Australia's total land area at

[4] https://eodatacube.eu/.

[5] https://cablab.readthedocs.io.

[6] https://www.opendatacube.org/.

a resolution of 25 square meters with more than 240,000 images that show how Australia's vegetation, land use, water movements, and urban expansion have changed over the past 30 years (NCI).

9.7 Conclusion

The advancement of Digital Earth drives collection of massive data such as transportation data, utility data, hazard data and forest data to monitor the Earth and support decision making. Valuable information extracted from the collected big data can further speed the development of Digital Earth. Digital Earth data continue to grow at a faster speed and with more heterogeneous types, leading to challenges to the lifecycle of data management including storage, processing, analytics, visualization, sharing, and integration. Fortunately, the emerging paradigm of cloud computing brings potential solutions to address these challenges. Compared with traditional computing mechanisms, cloud computing has the advantages of better data processing computing supports. The customizable configuration saves computing resources elastically, and data manipulation with higher security and flexibility offers secure data storage, transfer and sharing. Analytics enabled by cloud computing advance the process by allowing for automatic resource expansion when there are higher requirements.

To manage and analyze big Earth data, a service-oriented, scalable architecture based on cloud computing was introduced in a three-layer architecture: (1) the bottom data storage layer provides physical infrastructure, storage, and file systems; (2) the data query layer supplies data discovery capabilities with proper functionality and interoperability; and (3) the data processing layer supports extensibility, interoperability and scalability based on open source solutions and their variants from Earth science communities. With this architecture, big Earth data can be accessed and analyzed with low time latency or even in real time. The analysis results could be published by a web-based map server (e.g., GeoServer) or web-based notebook (e.g., Zeppelin) for visualization, public access, and collaboration, contributing to advancements in handling big data in Digital Earth to fulfill the requirements of scalability, extensibility and flexibility.

References

Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. Acm SIGMOD Rec 22(2):207–216

Amirebrahimi S, Rajabifard A, Mendis P et al (2016) A framework for a microscale flood damage assessment and visualization for a building using BIM–GIS integration. Int J Digit Earth 9(4):363–386

Anderson A (2015) Statistics for big data for dummies. John Wiley & Sons, Hoboken, NJ

Balakrishna C (2012) Enabling technologies for smart city services and applications. In: 2012 sixth International conference on next generation mobile applications, services and technologies. IEEE, Paris, France, 12–14 September 2012

Baumann P, Mazzetti P, Ungar J et al (2016) Big data analytics for earth sciences: the earthserver approach. Int J Digit Earth 9(1):3–29

Bereta K, Caumont H, Daniels U et al (2019) The copernicus app lab project: easy access to copernicus data. In: EDBT. pp 501–511

Big Data Virtualization (2019) https://www.techopedia.com/definition/29952/big-data-virtualization. Accessed 6 May 2019

Bizer C, Heath T, Berners-Lee T (2011) Linked data: the story so far. In: Amit S (ed) Semantic services, interoperability and web applications: emerging concepts. IGI Global, Hershey, PA, pp 205–227

Blachowski J (2016) Application of GIS spatial regression methods in assessment of land subsidence in complicated mining conditions: case study of the Walbrzych coal mine (SW Poland). Nat Hazards 84(2):997–1014

Boulos MNK, Al-Shorbaji NM (2014) On the internet of things, smart cities and the WHO Healthy Cities. Int J Health Geogr 13:10

Buck JB, Watkins N, LeFevre J et al (2011) SciHadoop: array-based query processing in Hadoop. In: Proceedings of 2011 International conference for high performance computing, networking, storage and analysis, ACM, New York, NY, 12–18 Nov 2011

Canada Line Vancouver Transit Map (2019) https://airfreshener.club/quotes/canada-line-vancouver-transit-map.html. Accessed 6 May 2019

Chang WL, Grady N (2015) NIST big data interoperability framework: volume 1, big data definitions (No. special publication (NIST SP)-1500-1).

Cudre-Mauroux P, Kimura H, Lim K-T et al (2009) A demonstration of SciDB: a science-oriented DBMS. Proc VLDB Endow 2(2):1534–1537

Dempsey C (2012) Where is the phrase "80% of data is geographic" from. https://www.gislounge.com/80-percent-data-is-geographic. Accessed 6 May 2019

Dillon T, Wu C, Chang E (2010) Cloud computing: issues and challenges. In: 2010 24th IEEE international conference on advanced information networking and applications, IEEE, Perth, Western Australia, 20–23 Apr 2010

Duffy DQ, Schnase JL, Thompson JH et al (2012) Preliminary evaluation of mapreduce for high-performance climate data analysis. NASA new technology report white paper

EarthCube (2016) EarthCube integration and testing environment(ECITE). https://www.earthcube.org/group/earthcube-integration-testing-environment-ecite. Accessed 6 May, 2019

EarthCube (2019) GeoSciCloud: Deploying Multi-Facility Cyberinfrastructure in Commercial and Private Cloud-based Systems https://www.earthcube.org/group/geoscicloud-deploying-multi-facility-cyberinfrastructure-commercial-private-cloud-based-systems. Accessed 6 May, 2019

EarthCube Brochure (2012) What is EarthCube? http://www.azgs.az.gov/images/agu-2012-earthcube-brochure-1.pdf. Accessed 11 Dec 2018

Eldawy A, Mokbel MF (2015) Spatial hadoop: a mapreduce framework for spatial data. In: 2015 IEEE 31st international conference on data engineering, IEEE, Seoul, South Korea, 13–17 Apr 2015

El-Mekawy M (2010) Integrating BIM and GIS for 3D city modelling: the case of IFC and CityGM. Doctoral Dissertation, KTH

Esch T, Uereyen S, Asamer H et al (2017) Earth observation-supported service platform for the development and provision of thematic information on the built environment—the TEP-Urban project. In: 2017 joint urban remote sensing event (JURSE), IEEE, Dubai, UAE, 6–8 Mar 2017

Firican G (2017) The 10 Vs of big data. Upside where data means business. https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx. Accessed 30 Jul 2019

Foster I, Zhao Y, Raicu I et al (2008) Cloud computing and grid computing 360-degree compared. arXiv preprint arXiv:0901.0131.

Friedman U (2012) Big data: a short history. https://foreignpolicy.com/2012/10/08/big-data-a-short-history. Accessed 6 May 2019
Gantz J, Reinsel D (2011) Extracting value from chaos. IDC iview 1142(2011):1–12
Geng Y, Huang X, Zhu M et al (2013) SciHive: array-based query processing with HiveQL. In: 2013 12th IEEE international conference on trust, security and privacy in computing and communications, IEEE, Melbourne, Australia, 16–18 Jul 2013
Geng Y, Huang X, Yang G (2014) Adaptive indexing for distributed array processing. In: 2014 IEEE international congress on big data, IEEE, Anchorage, AK, 27 June-2 Jul 2014
Giuliani G, Lacroix P, Guigoz Y et al (2017) Bringing GEOSS services into practice: a capacity building resource on spatial data infrastructures (SDI). Trans GIS 21(4):811–824
Gundersen E (2013) Visualizing 3 billion tweets. https://blog.mapbox.com/visualizing-3-billion-tweets-f6fc2aea03b0. Accessed 6 May 2019
Guo H (2017) Big earth data: a new frontier in earth and information sciences. Big Earth Data 1(1–2):4–20
Guo Z, Feng C-C (2018) Using multi-scale and hierarchical deep convolutional features for 3D semantic classification of TLS point clouds. Int J Geogr Inform Sci 1–20. https://doi.org/10.1080/13658816.2018.1552790
Guo H, Liu Z, Jiang H et al (2017) Big earth data: a new challenge and opportunity for digital earth's development. Int J Digit Earth 10(1):1–12
Hacker H, Trinitis C, Weidendorfer J et al (2010) Considering GPGPU for HPC centers: is it worth the effort? In: Keller R, Kramer D, Weiss J-P (eds) Facing the multicore-challenge: aspects of new paradigms and technologies in parallel computing. Springer, Berlin, Heidelberg, pp 118–130
Hamscher V, Schwiegelshohn U, Streit A et al (2000) Evaluation of job-scheduling strategies for grid computing. In: Buyya R, Baker M (eds) Grid computing—GRID 2000. Springer, Berlin, Heidelberg, pp 191–202
Hashem IAT, Chang V, Anuar NB et al (2016) The role of big data in smart city. Int J Inform Manag 36(5):748–758
Heuvelink GBM, Pebesma EJ (1999) Spatial aggregation and soil process modelling. Geoderma 89(1):47–65
Hong S-Y, O'Sullivan D (2012) Detecting ethnic residential clusters using an optimisation clustering method. Int J Geogr Inform Sci 26(8):1457–1477
Hu F, Xu M, Yang J et al (2018a) Evaluating the open source data containers for handling big geospatial raster data. ISPRS Int J Geoinform 7(4):144
Hu F, Yang C, Schnase JL et al (2018b) ClimateSpark: an in-memory distributed computing framework for big climate data analytics. Comput Geosci 115:154–166
Huffman GJ, Bolvin DT, Braithwaite D et al (2015) NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG). In: Algorithm theoretical basis document, version 4.1. NASA, Washington, DC
Hughes JN, Annex A, Eichelberger CN et al (2015) Geomesa: a distributed architecture for spatio-temporal fusion. In: Proceedings of SPIE 9473, geospatial informatics, fusion, and motion video analytics, SPIE, Washington, DC, 21 May 2015
Jiang Y, Li Y, Yang C et al (2018b) Towards intelligent geospatial data discovery: a machine learning framework for search ranking. Int J Digit Earth 11(9):956–971
Jiang Y, Li Y, Yang C et al (2018a) A smart web-based geospatial data discovery system with oceanographic data as an example. ISPRS Int J Geoinform 7(2):62
JoSEP, A. D., KAtz, R., KonWinSKi, A., Gunho, L. E. E., PAttERSon, D., & RABKin, A. (2010). A view of cloud computing. Communications of the ACM, 53(4).
JPL (2001) Izmit, Turkey 1999 earthquake interferogram. https://www.jpl.nasa.gov/spaceimages/details.php?id=PIA00557. Accessed 6 May 2019
Kofinas DT, Spyropoulou A, Laspidou CS (2018) A methodology for synthetic household water consumption data generation. Environ Model Softw 100:48–66
Krämer M, Senner I (2015) A modular software architecture for processing of big geospatial data in the cloud. Comput Graphics 49:69–81

Lahsen M (2016) Toward a sustainable future earth: challenges for a research agenda. Sci Technol Hum Values 41(5):876–898
Lam NSN (1983) Spatial interpolation methods: a review. Am Cartogr 10(2):129–150
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
Lenka RK, Barik RK, Gupta N et al (2016) Comparative analysis of SpatialHadoop and GeoSpark for geospatial big data analytics. In: 2016 2nd international conference on contemporary computing and informatics (IC3I), IEEE, Noida, India, 14–17 Dec 2016
Lepot M, Aubin J-B, Clemens HF (2017) Interpolation in time series: an introductive overview of existing methods, their performance criteria and uncertainty assessment. Water 9(10):796
Li W, Hsu C-Y (2018) Automated terrain feature identification from remote sensing imagery: a deep learning approach. Int J Geogr Inform Sci 1–24. https://doi.org/10.1080/13658816.2018.1542697
Li A, Yang X, Kandula S et al (2010) CloudCmp: comparing public cloud providers. In: Proceedings of the 10th ACM SIGCOMM conference on internet measurement, ACM, Melbourne, Australia, 1–30 Nov 2010
Li C, Wang S, Kang L et al (2014a) Trust evaluation model of cloud manufacturing service platform. Int J Adv Manuf Technol 75(1):489–501
Li W, Goodchild MF, Raskin R (2014b) Towards geospatial semantic search: exploiting latent semantic relations in geospatial data. Int J Digit Earth 7(1):17–37
Li Z, Yang C, Jin B et al (2015) Enabling big geoscience data analytics with a cloud-based, MapReduce-enabled and service-oriented workflow framework. PLoS ONE 10(3):e0116781
Li S, Dragicevic S, Castro FA et al (2016) Geospatial big data handling theory and methods: a review and research challenges. ISPRS J Photogramm Remote Sens 115:119–133
Li Z, Hu F, Schnase JL et al (2017a) A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce. Int J Geogr Inform Sci 31(1):17–35
Li Z, Yang C, Huang Q et al (2017b) Building model as a service to support geosciences. Comput Environ Urb Syst 61:141–152
Lipponen A (2017) New year's eve – average temperature at 18:00 local time. https://www.flickr.com/photos/150411108@N06/38663730944. Accessed 6 May 2019
Liu K, Yang C, Li W et al (2011) The GEOSS clearinghouse high performance search engine. In: 2011 19th international conference on geoinformatics, IEEE, Shanghai, China, 24–26 June 2011
Lui K, Yang C, Gui Z (2013) GeoSearch: a system utilizing ontology and knowledge reasoning to support geospatial data discovery. In: Workshop on semantics in geospatial architectures: applications and implementation, Pyle Center, University of Wisconsin-Madison, Madison, Wisconsin, 28–29 Oct 2013
Malik T (2014) GeoBase: indexing NetCDF files for large-scale data analysis. In: Wen-Chen H, Naima K (eds) Big data management, technologies, and applications. IGI Global, Hershey, PA, pp 295–313
Marr B (2015) Big data: using SMART big data, analytics and metrics to make better decisions and improve performance. John Wiley & Sons, Hoboken, NJ
Mayama K, Skulkittiyut W, Ando Y et al (2011) Proposal of object management system for applying to existing object storage furniture. In: 2011 IEEE/SICE international symposium on system integration (SII), IEEE, Kyoto, Japan, 20–22 Dec 2011
McCandless M, Hatcher E, Gospodnetic O (2010) Lucene in action: covers apache lucene 3.0. Manning Publications Co., New York, NY
Mc Cutchan M (2017) Linked data for a digital earth: spatial forecasting with next generation geographical data. In: International conference on spatial information theory, Springer, Cham, pp 91–96
Mell PM, Grance T (2011) Sp 800–145. The nist definition of cloud computing. NIST, Gaithersburg, MD
Microsoft (2011) Microsoft researchers' focus: eye on earth. https://www.microsoft.com/en-us/research/blog/microsoft-researchers-focus-eye-on-earth. Accessed 6 May 2019

NASA (2016) Getting petabytes to people: How the EOSDIS facilitates earth observing data discovery and use. https://earthdata.nasa.gov/getting-petabytes-to-people-how-the-eosdis-facilitates-earth-observing-data-discovery-and-use. Accessed 6 May 2019
Nativi S, Mazzetti P, Santoro M et al (2015) Big data challenges in building the global earth observation system of systems. Environ Model Softw 68:1–26
Nativi S, Mazzetti P, Craglia M (2017) A view-based model of data-cube to support big earth data systems interoperability. Big Earth Data 1(1–2):75–99
NCI (2019) Australian geoscience data cube. http://nci.org.au/services/virtual-laboratories/australian-geoscience-data-cube. Accessed 6 May 2019
NIST Big Data Public Working Group (2015) Nist big data interoperability framework. Use cases and general requirements. NIST, Maryland, US
NOAA (2017) GOES-16 first SEISS data.png. https://commons.wikimedia.org/wiki/File:GOES-16_first_SEISS_data.png. Accessed 6 May 2019
NSF (2009) Advisory committee for geosciences. http://www.nsf.gov/geo/acgeo/geovision/nsf_acgeo_vision_10_2009.pdf. Accessed 11 Dec 2018
NSF (2011) Earth cube guidance for the community. http://www.nsf.gov/pubs/2011/nsf11085/nsf11085.pdf.. Accessed 11 Dec 2018
Pagani GA, Trani L (2018) Data cube and cloud resources as platform for seamless geospatial computation. In: Proceedings of the 15th ACM international conference on computing frontiers, ACM, Ischia, Italy, 8–10 May 2018
Qi H, Gani A (2012) Research on mobile cloud computing: review, trend and perspectives. In: 2012 second international conference on digital information and communication technology and it's applications (DICTAP), IEEE, Bangkok, Thailand, 16–18 May 2012
Radke AM, Tseng MM (2015) Design considerations for building distributed supply chain management systems based on cloud computing. J Manuf Sci Eng 137(4):040906
Rahman MR, Lateh H (2017) Climate change in Bangladesh: a spatio-temporal analysis and simulation of recent temperature and rainfall data using GIS and time series analysis model. Theor Appl Climatol 128(1):27–41
Rathore P, Rao AS, Rajasegarar S et al (2017) Real-time urban microclimate analysis using internet of things. IEEE Internet Things J 5(2):500–511
Rew R, Davis G (1990) NetCDF: an interface for scientific data access. IEEE Comput Graph Appl 10(4):76–82
Rienecker MM, Suarez MJ, Gelaro R et al (2011) MERRA: NASA's modern-era retrospective analysis for research and applications. J Clim 24(14):3624–3648
Robert K (2000) Data flow controls water flow. https://www.spk.usace.army.mil/Media/Images/igphoto/2000748857. Accessed 6 May 2019
Robin A (2011) SWE CDM encoding standard, OGC. http://www.opengeospatial.org/standards/swecommon. Accessed 6 May 2019
Schabenberger O, Gotway CA (2017) Statistical methods for spatial data analysis. CRC Press, Boca Raton, FL
Segal B, Robertson L, Gagliardi F et al (2000) Grid computing: the European data grid project. In: 2000 IEEE nuclear science symposium. Conference record (Cat. No.00CH37149). IEEE, Lyon, France, 15–20 Oct 2000
Sharifzadeh M, Shahabi C (2004) Supporting spatial aggregation in sensor network databases. In: Proceedings of the 12th annual ACM international workshop on geographic information systems, ACM, New York, NY, 12–13 Nov 2004
Shi W, Cao J, Zhang Q et al (2016) Edge computing: vision and challenges. IEEE Internet Things J 3(5):637–646
Söderberg A, Dahlström P (2017) Turning smart water meter data into useful information: a case study on rental apartments in södertälje. Lund University, Lund, Sweden
Soille P, Burger A, Rodriguez D et al (2016) Towards a JRC earth observation data and processing platform. In: Proceedings of the conference on big data from space (BiDS'16), Publications Office of the European Union, Santa Cruz de Tenerife, 15–17 Mar 2016

Stefanakis E (2016) Discrete global grid systems—a new OGC standard emerges. GoGeomatics: Magazine of Gogeomatics Canada
Strobl P, Baumann P, Lewis A et al (2017) The six faces of the data cube. In: Proceedings of conference on big data from space (BiDS'17), Toulouse, France, 28–30 Nov 2017
USGS (2019) What is remote sensing and what is it used for? https://www.usgs.gov/faqs/what-remote-sensing-and-what-it-used?qt-news_science_products=7#qt-news_science_products. Accessed 6 May 2019
Vilches-Blázquez LM, Villazón-Terrazas B, Corcho O et al (2014) Integrating geographical information in the linked digital earth. Int J Digit Earth 7(7):554–575
Wang S, Liu Y (2009) Teragrid giscience gateway: bridging cyberinfrastructure and giscience. Int J Geographi Inf Sci 23(5):631–656
White T (2012) Hadoop: the definitive guide. O'Reilly Media, Inc., Sebastopol, CA
Xu C, Yang C, Li J et al (2011) A service visualization tool for spatial web portal. In: Proceedings of the 2nd international conference on computing for geospatial research & applications, ACM, New York, NY, 23–25 May 2011
Yang R (2016) A systematic classification investigation of rapid intensification of atlantic tropical cyclones with the ships database. Weather Forecast 31(2):495–513
Yang C, Raskin R, Goodchild M et al (2010) Geospatial cyberinfrastructure: past, present and future. Comp Environ Urban Sys 34(4):264–277
Yang C, Goodchild M, Huang Q et al (2011a) Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing? Int J Digit Earth 4(4):305–329
Yang R, Tang J, Sun D (2011b) Association rule data mining applications for atlantic tropical cyclone intensity changes. Weather Forecast 26(3):337–353
Yang C, Sun M, Liu K et al (2015a) Contemporary computing technologies for processing big spatiotemporal data. In: Kwan M-P, Richardson D, Wang D et al (eds) Space-time integration in geography and GIScience: research frontiers in the US and China. Springer Netherlands, Dordrecht, pp 327–351
Yang CP, Yu M, Sun M et al (2015b) Dasher cd: developing a data-oriented human-centric enterprise architecture for earthcube. In: AGU fall meeting abstracts. AGU, Washington, DC
Yang C, Huang Q, Li Z et al (2017b) Big data and cloud computing: innovation opportunities and challenges. Int J Digital Earth 10(1):13–53
Yang C, Yu M, Hu F et al (2017a) Utilizing cloud computing to address big geospatial data challenges. Comp EnvironUrban Syst 61:120–128
Yoo C, Ramirez L, Liuzzi J (2014) Big data analysis using modern statistical and machine learning methods in medicine. Int Neurourol J 18(2):50–57
Yu J, Wu J, Sarwat M (2015) Geospark: a cluster computing framework for processing large-scale spatial data. In: Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems, ACM, Seattle, US, 3–6 Nov 2015
Yu M, Yang C, Li Y (2018) Big data in natural disaster management: a review. Geosciences 8(5):165
Zaharia M, Chowdhury M, Das T et al (2012) Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of the 9th USENIX conference on networked systems design and implementation, USENIX Association, Berkeley, CA, 25–27 Apr 2012
Zhang Y, Kerle N (2008) Satellite remote sensing for near-real time data collection. In: Geospatial information technology for emergency response, 1st edn. CRC Press, Boca Raton, FL, pp 91–118
Zhang Y, Kunqing X, Xiujun M et al (2005) Spatial data cube: provides better support for spatial data mining. In: Proceedings 2005 IEEE international geoscience and remote sensing symposium. IGARSS'05, IEEE, Seoul, South Korea, 29–29 Jul 2005
Zhao H, Ai S, Lv Z et al (2010) Parallel accessing massive NetCDF data based on MapReduce. In: Wang FL, Gong Z, Luo X, Lei J (eds) International Conference on web information systems and mining. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 425–431
Zhao P, Qin K, Ye X et al (2017) A trajectory clustering approach based on decision graph and data field for detecting hotspots. Int J Geographi Inf Sci 31(6):1101–1127

Zhizhin M, Medvedev D, Mishin D et al (2011) Transparent data cube for spatiotemporal data mining and visualization. In: Fiore S, Aloisio G (eds) Grid and cloud database management. Springer, Berlin, Heidelberg, pp 307–330

Zikopoulos B, Barbas H. (2012). Pathways for emotions and attention converge on the thalamic reticular nucleus in primates. Journal of Neuroscience, 32(15), 5338–5350.

Yun Li is a Ph.D. candidate majoring in Earth Systems and Geoinformation Sciences at George Mason University. Her research mainly focuses on improving Earth data discovery using machine learning based methods, high performance computing, and spatiotemporal analytics for environmental and climate data.

Manzhu Yu is an Assistant Professor of GIScience at Pennsylvania State University. Her research focuses on spatiotemporal theories and applications, atmospheric modeling, environmental analytics, big data and cloud computing, and the ability of using the above to solve pressing issues in natural hazards and sustainability.

Mengchao Xu is a Ph.D. candidate in Earth Systems and Geoinformation Sciences program from George Mason University. His research mainly focuses on the cloud computing, high performance computing networks, distributed database systems, and precipitation downscaling. He manages a 500-nodes cluster specialized for cloud computing (supported by NASA, NSF and GMU).

Jingchao Yang is a Ph.D. student majoring in Earth Systems and Geoinformation Sciences at George Mason University. He has been contributing to several NSF and NASA funded projects at the NSF Spatiotemporal Innovation Center. His research mainly focuses on the smart city problems with machine- and deep-learning enabled spatiotemporal pattern analysis using multi-sensor (IoT) datasets.

Dexuan Sha is a Ph.D. student in Earth Systems and Geoinformation Sciences at the George Mason University. He researches arctic sea ice and melt pond from high spatial resolution aerial imagery with big spatial–temporal data mining and analysis methodology.

Qian Liu is a Ph.D. student in Geography and Geoinformation Science at George Mason University. Her research mainly focuses on the core areas of precipitation detection and analysis using AI-based methods, and outreaches to environmental and climate science and spatial temporal analytics.

Chaowei Yang is Professor of GIScience at George Mason University, where he established and directs the NSF spatiotemporal innovation center. His academic activities include spatiotemporal computing research, future GIScience leader education, and Geoinformatics outreach.