A COMPARATIVE ANALYSIS OF CONVENTIONAL HADOOP WITH PROPOSED CLOUD ENABLED HADOOP FRAMEWORK FOR SPATIAL BIG DATA PROCESSING

A. K. Tripathi [1,*], S. Agrawal [2], R. D. Gupta [3]

[1] GIS Cell, Motilal Nehru National Institute of Technology, Allahabad, Uttar Pradesh 211004, India - rgi1551@mnnit.ac.in
[2] GIS Cell, Motilal Nehru National Institute of Technology, Allahabad, Uttar Pradesh 211004, India - sonam@mnnit.ac.in
[3] Civil Engineering Department, Motilal Nehru National Institute of Technology, Allahabad -211004, India - rdg@mnnit.ac.in

Commission V, SS: Emerging Trends in Geoinformatics

KEY WORDS: Spatial Big Data, Cloud Computing, Hadoop, MapReduce

ABSTRACT:

The emergence of new tools and technologies to gather the information generate the problem of processing spatial big data. The solution of this problem requires new research, techniques, innovation and development. Spatial big data is categorized by the five V’s: volume, velocity, veracity, variety and value. Hadoop is a most widely used framework which address these problems. But it requires high performance computing resources to store and process such huge data. The emergence of cloud computing has provided, on demand, elastic, scalable and payment based computing resources to users to develop their own computing environment. The main objective of this paper is to develop a cloud enabled hadoop framework which combines cloud technology and high computing resources with the conventional hadoop framework to support the spatial big data solutions. The paper also compares the conventional hadoop framework and proposed cloud enabled hadoop framework. It is observed that the propose cloud enabled hadoop framework is much efficient to spatial big data processing than the current available solutions.

1. INTRODUCTION

Different forms of information gathering tools and techniques are available today. Organizations gather huge amount of data from different sources like sensors, remote sensing devices, scanners and digitizers (Berkovich and Liao, 2012). This type of huge information represents the big data (Bobade, 2014). These data are stored in either structured format like traditional dataset in form of rows and columns, or unstructured format like pdf document, images, videos, audios and email attachment. Organizations are burdened with large amount of unstructured data because 80 percent of the available data are unstructured and require significant storage space to manage them (Dhavapriya and Yasodha, 2016). The top international companies like Google, Yahoo and Facebook etc. use the big data processing tools to process huge amount of data collected at their server (Tiwarkhede and Kakde, 2015). For example, Facebook is handling large amount of data because in every second large number of clients can like the posts, upload or download the videos, comments on it and so on. Twitter stores around 250 million tweets per day and approximately 4 billion people are watching YouTube per day (Gao et al., 2014; Yang et al., 2016).

Today large amount of spatial data are generated which are utilized in different applications. It is estimated that 80-90% government data are interrelated to the spatial data (FGDC, 2017). In present time, various technologies like high speed internet, low cost smart devices and GPS has increased the production and accessibility of spatial data (Dhavapriya and Yasodha, 2016). GIS is also dealing with the huge volume of spatial datasets which are collected from various sources like LiDAR point cloud, remote sensing data and satellite images (Goodchild, 2007; Yang et al., 2011). So the proper tools are required to analyze and process these large data and extract meaningful information from big data.

Processing of spatial big data is always been a complex and time consuming process. It cannot be processed and analyzed by the traditional tools and techniques such as relational database and desktop tools. Spatial big data cannot be stored at one single machine. It is usually stored in a distributed manner on multiple computers. Several distributed High Performance Computing (HPC) models like Graphics Processing Units (GPUs) (Owens et al., 2006; Tang and Feng, 2017) and MapReduce (Gao et al., 2014; Giachetta, 2015) are available to handle big data. Among them, MapReduce is most widely accepted models because it works on the concept of parallel computing. This technology provides high throughput, low operational costs, minimum processing time and reduce risk (Bhosale and Gadekar, 2014; Shilpa and Kaur, 2013).

The main objective of the paper is to present the cloud enabled hadoop framework based on open source cloud provider. OpenStack is used as a cloud service provider which is an open source software to develop the cloud environment. Moreover, the paper also presents the comparative analysis of conventional hadoop with proposed cloud enabled hadoop framework for big spatial data processing.

2. BACKGROUND

2.1 Spatial big data

Big data represents the flood of digital form of data collected from different sources. Big Data refers to the voluminous amount data present in structured, semi-structured or unstructured format that need to be analysed to extract useful information from them. These data are characterized on the basis of five V’s (Lee and Kang, 2015). They are discussed below in the reference of spatial big data.

[*] Corresponding author

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-5-425-2018 | © Authors 2018. CC BY 4.0 License.

- Volume: It represents the amount of data. Currently, data stored in organizations grown from gigabytes (GB) and terabytes (TB) to petabytes (PB) (Yang et al., 2016). Different spatial data sources can produce data in GB or TB per hour (Dasgupta, 2013). So the system which processes such huge amount of spatial data should be scalable and distributed. Big data architecture given in this paper provides infrastructure which can easily handle it.
- Velocity: It represents the data processing speed. Today many of the sensors produce huge data in a very short amount of time. This requires faster data processing rate so that user can get the processed information in real time.
- Variety: It refers to the different forms of data and sources. These spatial data collected from different traditional and conventional mechanism like remote sensing, photogrammetry, surveying, volunteer geographic information, Global Positioning System (GPS) and geo-tagging (Tang et al., 2017). Spatial data are mostly heterogeneous. It is necessary to combine such heterogeneous form of spatial data on the basis of same data exchange format, accuracies, scale and reference system. It requires efficient algorithms to structure, index and manage such data like NoSQL and hadoop.
- Veracity: Much of the data are gathered from unverified sources with low accuracy and quality. It varies from data to data based on the data sources. So it requires proper quality assessment mechanism to ensure quality and accuracy.
- Value: It is categories on the basis of quality, higher resolution, admissible cost and user satisfaction. Adoption of conventional technologies like cloud computing, MapReduce ensures such security protocol.

2.2 Comparison table of Big Data and Spatial Big Data

A comparison between the big data and spatial big data is performed which is summarized in table1.

	Big Data	Spatial Big Data
Raster	Photographs, graphics, medical records such as X-rays, ECG and MRI images	Time series maps, global topographic data, country, national, region wise orthophotos, etc.
Vector	2D and 3D graphics	Cadastral data, network data, environmental data
Point cloud	3D objects such as medical, art, geology, robotics, etc.	LiDAR data
Text based	Text messages, sensors text data, social network, comments	Geolocation and geosocial based text data (coordinates, address, etc.)
Storage format	Relational DBMS, distributed file system, key-value store	Array database, distributed file system, key-value store

Table 1. Comparison of Big Data and Spatial Big Data

2.3 Key Challenges of Spatial Big Data

From the overview of the spatial big data, it can be found that analysis, processing and extraction of useful information from big data undergoes through a numerous challenges. Some of them are listed below

- Data storage: Storage of the big data is the crucial problem from the several years. There are number of traditional physical storage devices but they are not efficient to store such huge amount of data (Krämer and Senner, 2015). Velocity of the data, i.e. the scalability of quick response of processing of spatial big data, directly depends upon the volume of the data which depends upon the availability of the storage device.
- Data processing: Processing of the large data volume is not easy by the traditional computing systems. It requires high computing resources i.e. CPU, network and storage that can be scaled up on the basis the data volume and processing (Labrinidis and Jagadish, 2012). Further, the processing of big data is also not easy by the traditional processing algorithms. For example, the real time data such as climate data, sensors data, etc. can't processed by the traditional algorithms. It requires highly efficient and effective reduction algorithms to remove irrelevant, noisy data from these big datasets (Zhai, Ong, and Tsang 2014).
- Data visualization: Data visualization is the critical part of spatial big data, as it covers the hidden pattern of the datasets. It contains the heterogeneous (structured and unstructured) datasets (Padgavankar and Gupta, 2014). Data visualization covers the several parameters like interactive graphical user interface, web based interface for filtering and extracting datasets, integrated and intuitive visual pattern of data. Designing of these functionalities is a challenging task because different features of the big data include the multiple data sources, high spatial resolution and high dimension of spatial data.
- Data integration: It is the crucial part of data management. In the spatial domain, the spatial data integration is the key for data analysis and decision making process. There are several challenges like schema mapping, data fusion and record linkage which are barrier to the spatial data integration. Further, the metadata is essential to perform such mapping and data fusion to extract the crucial information. However, automatic metadata creation from spatial big data is a challenging task.
- Data transmission: Data transmission is essential component of the spatial big data in the every stage of the spatial data life cycle. It is needed in data collection, data integration from multiple sources, transfer of integrated data to the processing platform and data hosting. Therefore, efficient compression techniques are required for effective data transmission (Yang et al., 2016).

2.4 Distributed Computing

Distributed computing refers to to a set of interconnected computers, which communicate and coordinate between each other through the message passing. All the computers have capability to interact each other for storage sharing and computation to achieve a common objective (Ali and Khan, 2015). Distributed computing is heterogeneous in nature that's why all the interconnected nodes have different processing speed, network topology, operating system and communication medium (Chhabra et al., 2006; Gu et al., 2005). As the hundreds of computers are connected to build the distributed computing environment, it is necessary to distribute the work load on the computing nodes to gain the maximum throughput, efficiency, speedup and system utilization (Ali and Khan, 2012). Massive computational speed and storage capacity have been gained due the distributed structure. All the interconnected nodes have features like no common physical clock, no shared memory, geographical separation, autonomy and heterogeneity (Olasz et al., 2017). Distributed computing environment is different from the parallel computing in the reference of memory sharing. No memory is shared between the nodes in distributed computing to execute a job. Nodes are communicated through message queues. There are mainly two popular distributed architectures that are presented henceforth.

a) Client-Server model

In this architecture, client makes request for the service provided by the server as shown in figure 1. Clients initiate the scheduling of jobs to the server, which process that job and send result back to the client. All the data are stored in the centralized server. The main focus of the model is to share the information.

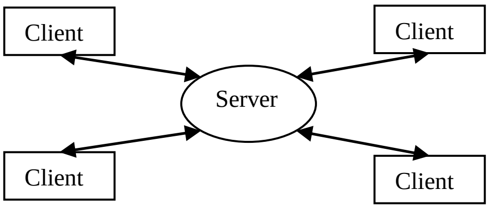

Figure 1. Client-server model

b) Peer to Peer model:

In the peer to peer network, all the nodes are interlinked to each other as shown in figure 2. There is no centralized control as in client server model. In peer to peer network, all the nodes have its own storage capability, bandwidth and computation speed. There is no difference between client and server. Any node can act as client or server based upon the requirement.

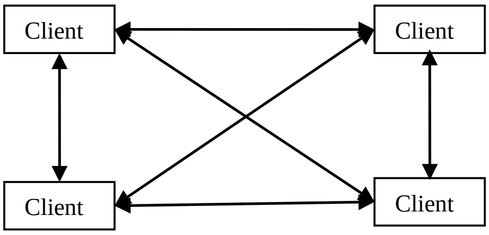

Figure 2. Peer to peer model

2.5 Cloud Computing

With the development of internet, information sharing became easy. Next step is the usability of this technology to share the computing resources over the web. Cloud computing provides this type of functionality. Cloud computing refers to the sharing of computing resources over the internet, instead of using local computer (Yang et al., 2016). Cloud computing has adopted several conventional computing mechanism like grid computing, utility computing. Cloud provides the elastic virtualized computing resources on the demand basis and pay per use model. Virtualization refers the abstraction of the physical resources by providing virtualized computing resources (Zhang et al., 2010).

Cloud computing is a promising approach to deliver storage and computing environment to the datacenter, which is accessed from anywhere in the world. Several public cloud service providers are available now-a-days, for example, Amazon EC2, Microsoft Azure, IBM, Google App Engine (GAE) etc. Amazon EC2 is most widely used cloud service provider which provides virtual computing environment for the user to develop their application. It provides Amazon EC2 for the computation purpose and Amazon S3 to store the data (Gao et al., 2014). Cloud computing have many advantages such as low-cost, scalability, reliability, security, quality of service and flexibility.

3. HADOOP: BIG DATA STORAGE AND PROCESSING

Hadoop framework is designed to provide scalable, reliable and shared storage infrastructure for the user community. Hadoop is an Apache open source framework written in Java that allows distributed computing environment for processing of huge amount of data across cluster of computers in parallel manner (Kumari, 2014). It is designed to scale up a single server to multiple machines and each machine act as a separate local computational and storage space. Hadoop has two main components: Hadoop Distributed File System (HDFS) and MapReduce framework. HDFS supports the storage of the big data within distributed environment while the MapReduce provide the processing of the information ((Bhosale and Gadekar, 2014; Kaur and Kaur, 2015).

HDFS is going to distribute the large file into the blocks and each block is associated with the each node of the hadoop cluster system (Kaur and Kaur, 2015). Hadoop provides the fault tolerant system by the replication of the data i.e. clusters have three copies of the blocks stored in cluster. It is very important because even if one of node fails, the system will still work properly. It searches the copy of the block into other node. Hadoop architecture has the following components as shown in figure 3. It works on the concept of master and slave node. Masters have the name node, secondary node and Job Tracker, while slaves have the data node and Task Tracker. Each component within master and slaves are associated to each other, while the name node accomplish with the data node. Job Tracker controls the task associated with the Task Tracker. The blocks are stored in the data node and it is the responsibility of the name node to notify that which block is attached to with which data node. Name node also handles the file in data node. Although there is strong relationship between the name node and data node, still they work as "loosely coupled" manner.

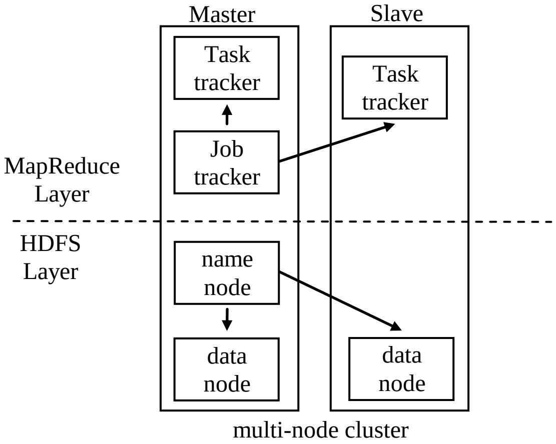

Figure 3. Hadoop Architecture (Bhosale and Gadekar, 2014)

Now the next important component of the hadoop architecture is MapReduce (Bhosale and Gadekar, 2014; Bobade, 2014; Dhavapriya and Yasodha, 2016). It is a programming module that processes and generates the large data sets attached with the clusters. It is two phase process: 'map' and 'reduce' phase. When the programmer write the code to access the data from the cluster it first send it to Job Tracker that assign task to the Task Tracker node if it is idle by using map function. The input datasets are divided into the multiple chunks of datasets, each of which is assigned a map task that run in parallel manner based on the key and value pair and produces a transformed set of key and value pair of output. Further framework shuffle and sort the transformed output and transmit the intermediate key and value pair to the reduce tasks which group them into the final result. There is a major drawback of the hadoop framework, i.e., when the master node fails, the whole system is shutdown. Therefore there is a single point failure in hadoop architecture and that’s why high computational power hardware is needed in master nodes.

4. PROPOSED CLOUD ENABLED HADOOP FRAMEWORK

The emergence of cloud benefits various areas of applications including the spatial community. As per discussion about the challenges of the spatial big data, it can be easily elaborated that cloud can provide the better infrastructure to handle spatial big data. Cloud enabled hadoop framework has been proposed in this paper. It is shown in figure 4. This framework presents the implementation of conventional hadoop framework in cloud environment. The proposed framework uses the open source OpenStack software to develop the cloud framework. Openstack provides the Infrastructure as a Service (IaaS) to the user to create the Virtual Machines (VMs). These VMs are runs by hypervisiors such as KVM or Xen. This framework presents the multilevel architecture. It’s steps are discussed below:

1. Resource Provisioning: First step follow the resource provisioning through the OpenStack. To setup hadoop cluster manually, cloud provides the VMs dynamically on demand basis. Based upon the amount of the spatial big data, user can get the access of VMs automatically.

2. Spatial Data Management and Sharing: In this framework, large spatial data is stored into the cloud HDFS. Data stored in cloud HDFS is distributed into the cluster of VMs. Efficient spatial data management and sharing is important, especially for processing such huge amount of data.

3. Data Partitioning and Mapping: This step performs the MapReduce operation. Large size spatial datasets are distributed into cloud data nodes. To distribute the data over the data nodes, proper indexing is required which is further utilized for map tasks. As spatial data consists of the vector and raster data format, it is necessary to perform localized geometry partitioning algorithm. Many popular indexing mechanisms are available including R*-tree, k-d tree, quadtree and hash based indexing which can easily adopt the key-value pair pattern for mapping functionality. Moreover, Wang et al. (2016) presented the Hilbert R+ tree indexing mechanism for spatial big data processing. These partitioning algorithms are applied based upon the category of spatial information. Map operation is done using the key-value pairs and performs locally. All these chuck files are processed concurrently using mapper.

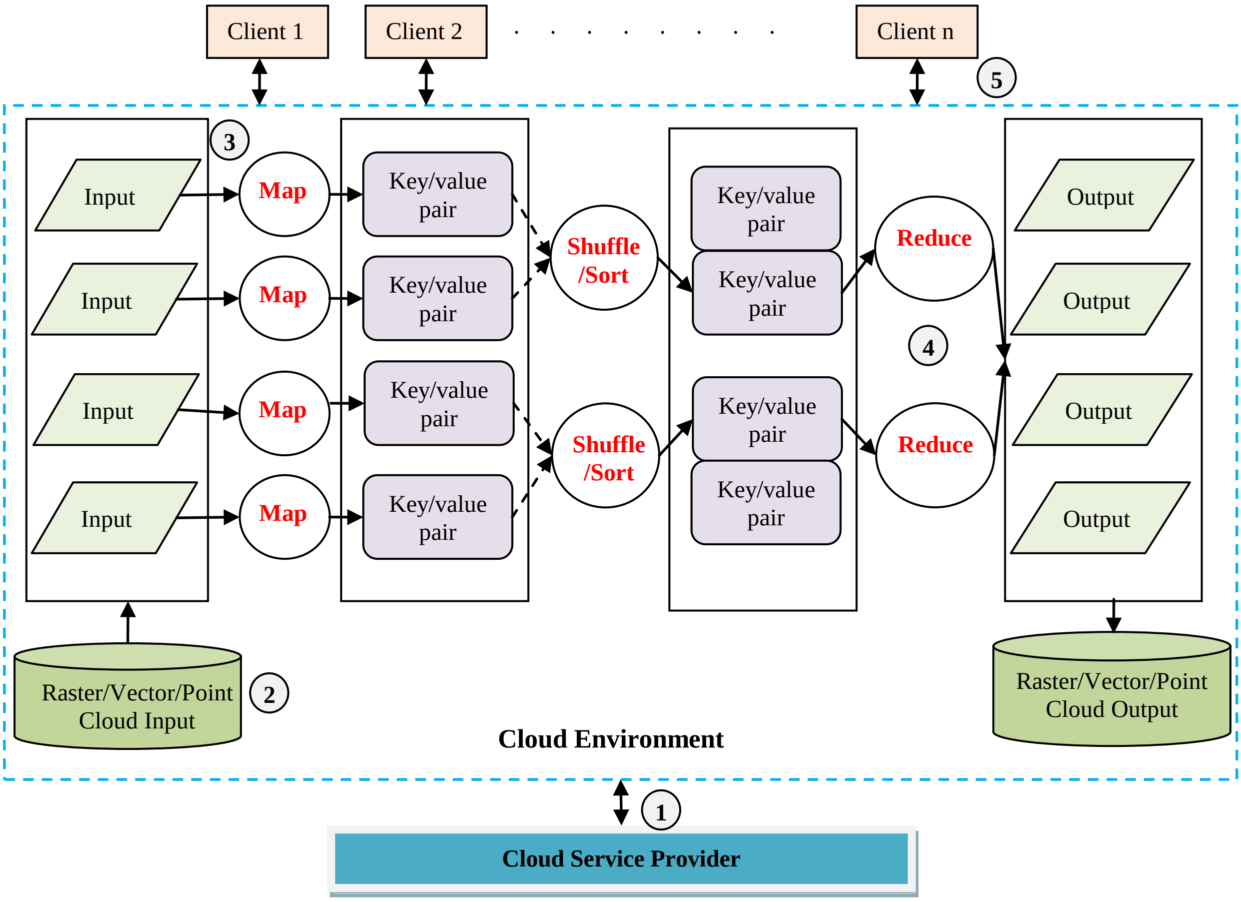

Figure 4. Cloud enabled hadoop framework

4. Shuffle and Reduce: The reduce operation is performed for post processing which shuffle the whole results of map operations. If post processing is not required, then the reducer directly merges all the map operation results based on intermediate values associated with the same key. If the size of final processed spatial data is greater than the allocated block size, the repartitioning is performed again. Such repartitioning is performed by the additional MapReduce process.

5. Querying and Processing: This step enables clients to perform query on such distributed computing environment. Cloud enabled hadoop infrastructure make client capable to access wide variety of spatial data. Such framework provides various key functionalities like authentication, management and accountability.

5. COMPARISON BETWEEN CONVENTIONAL AND CLOUD BASED HADOOP FRAMEWORK

The comparison of conventional hadoop framework with the prosed framework is presented in Table 2.

	Conventional Hadoop	**Cloud enabled Hadoop**
Infrastructure setup	It uses thousands of physical machines to setup the distributed computing environment.	It uses VMs to setup the distributed computing environment which is comparatively easier.
Cost efficiency	High processing computing systems are purchased, therefore cost is high.	High processing computing systems are rented on demand and paid as per use, therefore cost is less.
Scalability	Scalability is the major issue	On demand, resources provisioning of cloud makes it scalable.
Internet Connectivity	The management and configuration of resources are localized therefore continuous internet connectivity is not required.	Computing resources are present at different physical locations therefore continuous internet connectivity is required.
Security and Privacy	It provides the localized storage of data. Therefore it is more secure.	This is a major issue as the data is stored at the cloud service provider end.

Table 2. Comparison table of conventional hadoop framework with the prosed framework

6. CONCLUSIONS

The present paper covers the big data and spatial big data and their processing. Cloud enabled hadoop framework is proposed in the paper which can facilitate the distributed environment to process spatial big data. The main focus of this paper is to use the high performance computing technology and software to address the problem of spatial big data. The comparison of the conventional hadoop framework with the proposed framework is also covered in the paper. From the comparison, it is clearly identified that the use of cloud computing provides better distributed infrastructure to handle spatial big data. On demand resource provisioning makes user more capable. However, the challenges of spatial big data also helps to understand the need of utilizing cloud computing. Future spatial applications, such as fog forecasting, disaster assessment and the smart city projects, produce huge amount of spatial information, thus requiring real time computation and processing techniques. The proposed framework is very useful to manage and process such huge spatial data and meet the future requirements.

REFERENCES

Ali, M.F., Khan, R.Z., 2015. Distributed Computing: An Overview. International Journal Advanced Networking and Applications 7, 2630–2635.

Ali, M.F., Khan, R.Z., 2012. The study on load balancing strategies in distributed computing system. International Journal of Computer Science & Engineering Survey (IJCSES) 3, 19–30.

Berkovich, S., Liao, D., 2012. On Clusterization of " Big Data " Streams, in: Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications. ACM, Washington, D.C., USA, p. 1-6. https://doi.org/10.1145/2345316.2345347

Bhosale, H.S., Gadekar, D.P., 2014. A Review Paper on Big Data and Hadoop. International Journal of Scientific and Research Publications 4, 2250–3153.

Bobade, V.B., 2014. Survey Paper on Big Data and Hadoop. International Research Journal of Engineering and Technology (IRJET) 3, 585–590.

Chhabra, A., Singh, G., Waraich, S.S., Sidhu, B., Kumar, G., 2006. Qualitative Parametric Comparison of Load Balancing Algorithms in Parallel and Distributed Computing Environment, in: Word Academy of Science, Engineering and Technology. pp. 39–42.

Dasgupta, A., 2013. BIG DATA : The Future Is Now. Geospatial World.

Dhavapriya, M., Yasodha, N., 2016. Big Data Analytics: Challenges and Solutions Using Hadoop, Map Reduce and Big Table. International Journal of Computer Science Trends and Technology (IJCST) 4, 5–14.

FGDC, 2017. Homeland Security and Geographic Information Systems-How GIS and mapping technology can save lives and protect property in post-September 11th America [WWW Document]. Public Health GIS News and Information. URL https://www.fgdc.gov/resources/whitepapers-reports/white-papers/homeland-security-gis

Gao, S., Li, L., Li, W., Janowicz, K., Zhang, Y., 2014. Constructing gazetteers from volunteered Big Geo-Data based on Hadoop. Computers , Environment and Urban Systems 61, 172–186. https://doi.org/10.1016/j.compenvurbsys.2014.02.004

Giachetta, R., 2015. A framework for processing large scale geospatial and remote sensing data in MapReduce environment. Computers and Graphics 49, 37–46. https://doi.org/10.1016/j.cag.2015.03.003

Goodchild, M.F., 2007. Citizens as sensors : the world of volunteered geography. GeoJournal 69, 211–221. https://doi.org/10.1007/s10708-007-9111-y

Gu, D., Yang, L., Welch, L.R., 2005. A Predictive , Decentralized Load Balancing Approach, in: 19th IEEE International Parallel and Distributed Processing Symposium. IEEE, Denver, CO, USA, p. 8. https://doi.org/10.1109/IPDPS.2005.60

Kaur, G., Kaur, M., 2015. Review Paper On Big Data Using Hadoop. International Journal of Computer Engineering & Technology (IJCET) 6, 65–71.

Kumari, S., 2014. A Review Paper on Big Data and Hadoop. International Journal of Scientific and Research Publications 4, 2250–3153.

Lee, J., Kang, M., 2015. Geospatial Big Data : Challenges and Opportunities. Big Data Research 2, 74–81. https://doi.org/10.1016/j.bdr.2015.01.003

Olasz, A., Thai, B.N., Kristóf, D., 2017. Development of a new framework for Distributed Processing of Geospatial Big Data. International Journal of Spatial Data Infrastructures Research 12, 85–111.

Owens, J.D., Luebke, D., Govindraju, N., Harris, M., Kruger, J., Lefohn, A.E., Purcell, T.J., 2006. A Survey of General Purpose Computation on Graphics Hardware. Computer Graphics Forum 26, 80–113. https://doi.org/10.1111/j.1467-8659.2007.01012.x

Shilpa, Kaur, M., 2013. BIG Data and Methodology-A review. International Journal of Advanced Research in Computer Science and Software Engineering 3, 2277–128.

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

Tiwarkhede, A.S., Kakde, V., 2015. A Review paper on BIG Data Analytics. International Research Journal of Engineering and Technology 4, 2395–56.

Yang, C., Goodchild, M., Huang, Q., Nebert, D., Raskin, R., Xu, Y., Bambacus, M., Fay, D., 2011. Spatial cloud computing : how can the geospatial sciences use and help shape cloud computing ? International Journal of Digital Earth 4, 305–329. https://doi.org/10.1080/17538947.2011.587547

Yang, C., Huang, Q., Li, Z., Liu, K., Hu, F., 2016. Big Data and cloud computing : innovation opportunities and challenges. International Journal of Digital Earth 10, 13–53. https://doi.org/10.1080/17538947.2016.1239771

Zhang, Q., Cheng, L., Boutaba, R., 2010. Cloud computing: State-of-the-art and research challenges. Journal of Internet Services and Applications 1, 7–18. https://doi.org/10.1007/s13174-010-0007-6