ANALYTICS AND STORAGE OF BIG DATA

Shubham Upadhyay, Rakesh Manwani, Saksham Varshney and Sarika Jain

NIT Kurukshetra, Kurukshetra, Haryana, 136119, India

Abstract

Data generated by the devices and the users in modern times is high in volume and variable in structure. Collectively termed as Big Data, it is difficult to store and process using traditional processing tools. Traditional systems store data on physical servers or cloud resulting in higher cost and space complexity. In this paper, we provide a survey of various state-of-the-art research works done to handle the inefficient storage problem of Big Data. We have provided comparative literature to compare existing works to handle Big Data. As a solution to the problem encountered, we propose to split the Big Data into small chunks and provide each chunk to a different cluster for removing the redundant data and compressing it. Once every cluster has completed its task, the data chunks are combined back and stored on the cloud as compared to physical servers. This effectively reduces storage space and achieves parallel processing, thereby decreasing the processing time for very large data sets.

Keywords

Big Data, Cloud Computing, Data Analytics, Data Compression, Storage System,

1. Introduction

Digitization is generating a huge amount of data, and Information Technology Organizations have to deal with this huge data [1], which is very difficult to manage and store. This data comes from various sources like social media, IoT devices, sensors, mobile networks, etc. According to some figures, 2/3rd of the total data has been generated in the last 2 years [2]. According to Intel, smart cars generate 4000GB of data per day.

Traditionally, companies prefer to deploy their servers for data storage, but as the volume of data increases it becomes challenging for the companies to manage the infrastructure required and the cost associated with it, this also poses flexibility issues. The problem related to the management of data can be handled by using infrastructures like Cloud, which provide close to unlimited storage along with services such as data security because of which data owners don't have to put much effort into it and can focus on their day-to-day tasks.

Along with being large, the data is also complex which possesses problems when it has to be processed with traditional processing tools. For this, we need some dedicated tools which will facilitate the processing of this data, which are part of Big Data Computing. It involves master-slave architecture in which there is a single master node that assigns a task to slave nodes, which works in a parallel fashion. It facilitates faster processing. The data is generally scattered and broadly classified into three types:

- Structured: Data that can be stored in tables, made of rows and columns comprising a database is termed as structured data. This type of data can be easily processed. Eg. Relational data.
- Semi-Structured: Similarly, data that cannot be stored in form of a database but can be easily analyzed is termed semi-structured data. They usually occupy less space. Eg. XML data.
- Unstructured data: Unstructured data have an alternative platform for storage and management that is mainly used in an organization with business intelligence. Eg. Pdf, Media files.

As each structure has different features, they are needed to be processed by different tools and hence, making it difficult to define a single mechanism to process big data efficiently. Along with complex structure, big data is also characterized by 5 V's which defines the things a system developer has to keep in mind while dealing with Big Data. These V's are:

- Velocity: Velocity is the speed of data generation, analysis, and collection. With each day, the velocity of data keeps on increasing.
- Volume: Volume is the amount of data, which is generated from social media, credit cards, sensors, etc. The volume of data is so large that it is difficult to manage, store, and analyze it.
- Value: In the current scenario, data is money. It's worthless if we can't extract value from it. Having huge data is something, but if we can't extract value from it, then it is useless.

International Semantic Intelligence Conference (ISIC 2021)

✉ supadhyay567@gmail.com (S. Upadhyay); rehan.manwani.56@gmail.com (R. Manwani); sakshamvarshney8@gmail.com (S. Varshney); jasarika@gmail.com (S. Jain)
iD 9729467453 (S. Upadhyay)

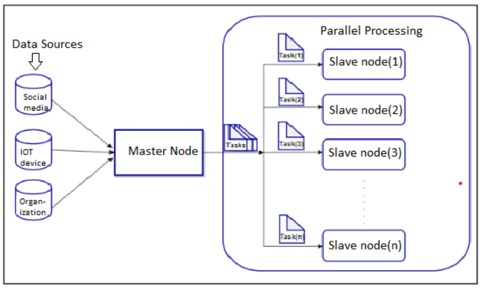

Figure 1: Parallel Processing Environment.

- Variety: Data is available in three forms that are structured, semi-structured, and unstructured. The volume of structured data is very less as compared to unstructured data.

- Veracity: Veracity refers to the quality of the data. It means how accurate data is. It tests the reliability and accuracy of the content.

Cloud is a better storage platform where our data is easily accessible and secure. So business firms have started storing their data in the cloud. But the rate of growth of data is exponential; as a result, cloud servers also lack such a huge volume of storage. Therefore, there emerges a need to select important data and store it in a way that it could fit in less memory space and it should be cost-effective. Now to achieve this objective we require a system that can perform this task in less time but the single system is not able to do this task efficiently thus we require an environment where we can achieve parallel processing to perform this task fast.

Fig. 1 shows a way to process it faster by dividing data into small chunks and assigning each chunk to a cluster for processing the data chunk provided by the master node. This will achieve parallel processing and increases the rate of processing.

Challenges with Big Data Storage System

Where there are opportunities, there are challenges. With various benefits that can be gained with large data sets, big data storage possesses various challenges too. Various computational methods that work well with small data sets won't work well with big data sets. The following are some major challenges with big data storage:

- **Format of Data:** While working for Big Data Storage Management one of the prime challenge is to make the system which will deal with both structured and unstructured data, equally.

- **Data Storage:** Volume, Variety, and Velocity of Big Data lead to various storage challenges. Traditional Big Data Storage is quite challenging as Hard Disk Drives fail (HDDs) very often and can't assure efficient data storage with several data protection mechanisms. Adding to it, the velocity of big data generates a need for scalable storage management to cope up with it. Though, Cloud provides a potential solution to address the problem with unlimited storage options which highly faults tolerant, transferring the Big Data to and hosting on the cloud is quite expensive with this huge amount of data [3].

- **Data Transmission:** Transmission of data consists of several stages like a) Collecting data coming from different origins, like sensors, social media, etc. b) Integration of the collected data that comes from different origins. c) the Controlled transfer of integrated data to processing platforms. d) Movement of data from server to host [4]. This amount of transferring data within different stages is quite challenging in various manners.

- **Data Processing:** Preparing enormous volumes of information requires devoted computing assets and this is mostly taken care of by the expanding pace of CPU, network, and capacity. Anyway, the computing assets required for handling huge information far surpassed the preparing power offered by customary driving ideal models.

- **Data Duplication:** In a big data environment, most of the data sets have identical content and are redundant. Data duplication increases the cost and also takes more space.

Benefits of Data Analytics and Compression

Data Analytics comprises collecting data, removing data redundancy, analyzing results, etc. and compression comprises reducing the file size to save space. These both combined can be used to benefit our research question. The following are the benefits of data analytics and compression of data:

- **Less disk space:** As the size of big data is very large, compressing it after analyzing can help in reducing disk space to a great extent. This process releases a lot of space on the drive and as a result memory results are closed up, which reduces the time that is required to retrieve the data from a server.

- **Faster file transfer:** After the file compression, the size of the file will be reduced. So time for transmitting a file with a reduced size will be faster.

- **Accessibility:** Accessing managed data is relatively easier as it allows a faster searching from hugely populated data. Also, data can be accessed remotely from any place with internet connectivity.

- **Cost Reduction:** Using the cloud as a storage medium helps in reducing hardware costs as well as cutoff energy costs. We can rent as much space as we want at a minimal cost.

- **Virtualization:** Cloud provides backup for big data and takes off the burden of growing data from the enterprises. To provide a backup, it is recommended to make virtual copies of the applications instead of making the physical copies of the analytics.

- **Scalability:** Big data management allows the application to grow exponentially as it deals with the storage space by itself. It reduces the need for new servers or supporting hardware as it manages the data in the existing ones.

A lot of work has been done in the concerned direction by different authors, and the following are the contributions of this work: (1) An exhaustive study of the existing systems has been done and based on three parameters, namely data processing tools used, data compression technique, and data storage option used, a review has been done and summarized to get a gist of all the existing ways to deal with Big Data. (2) Based on the comparative study mentioned above few gaps in the existing system are extracted and a solution to fill those gaps is discussed. This paper is structured into four sections. After introducing the paper in section 1, we move to section 2 discussing the works done so far in the concerned direction and a comparative study of systems of different authors that deal with big data. Section 3 provides the gaps that come out of the previously done work. And a solution is proposed to fill these gaps. Finally, section 4 concludes the work, and targeted future work is also mentioned.

2. Related Work

Various systems tried to achieve an efficient data storage system and tried various techniques. Based on some differentiating parameters, we have done a comparative study between different Storage Management Systems proposed by different authors in their research work. So, to draw a better technique between various systems we will use the following differentiating parameters (a) Data Processing, (b) Data Compression, and (c) Data Storage. Different data processing/ pre-processing techniques are used by various systems to analyze the big data and reduce the data redundancy and we will differentiate the existing systems based on these techniques also. The following are some important data processing tools for big data:

- **Apache Hive:** Hive is a software project developed by Apache. It is used for providing query and analysis of Big Data. It has a SQL like interface.

- **Apache Spark:** Spark is an open-source big data analytical engine that provides an environment for cluster computing and parallel processing of big data. It comes with inbuilt modules of SQL, machine learning, graph processing, etc.

- **Hadoop Map Reduce Another Tool:** that can be used for programming the model of Big Data is Hadoop Map Reduce. To write the programs for Map Reduce various languages like Java, Python, C++ are used popularly.

- **Apache Giraph:** Apache Giraph is real-time graph processing software built for high scalability which is used to analyze social media data. Many multinational companies are using Giraph, tweaking the software for their purposes.

- **PySpark:** PySpark is a python API processing on the programming model of Spark to Python. It is a necessary supplement API when one wants to process Big Data and apply machine learning algorithms at the same time.

- **Prophet Method:** Prophet Prophet is a procedure for foreseeing time game plan data reliant on an additional substance model where examples are fit annually, step by step, and ordinary consistency, notwithstanding event results. It's best results are with the time course of action having strong ordinary effects and a couple of times of recorded data.

- **Neural Network:** Neural frameworks are a ton of computations, shown openly after a man's cerebrum, that is expected to see plans. They unravel material data through such a machine acknowledgment, naming, or grouping unrefined information. Neural frameworks or connectionist structures are preparing systems questionably spurred by the natural neural frameworks that set up human cerebrums. Such structures ”learn” to perform tasks by pondering models, generally without being altered with task-unequivocal guidelines.

- **MongoDB:** MongoDB is an intermediate database management system between key-value and traditional Relational Database Management System. It is a document-oriented DB system and is classified as a NoSQL database system and is the database for Big Data processing.

- **NoSQL Database:** It stands for “Not Only SQL” and is used against RDBMS where we used to build the schema before the actual database and the data is stored in the form of tables.

- **Hashing:** A hash work is any capacity that can be utilized to delineate subjective size to fixed-size qualities. The features that are used to fill a table of fixed-size, termed as a hash table. The consequence of a hash work is known as a hash value or basically, a hash.
- **Docker:** DOCKER is one of the top organizations on the planet which is a light computing framework that provides containers for service. These containers are open source containers and are easily accessible to anyone free of cost. It is because of this only that container administrations are enjoying enormous interest. The Docker gives the containers progressively secure and easy to use benefits. Because it gives customary updates to the containers which is the reason the holder won't bargain with the speed of its execution.

Data Compression techniques are also used by the systems as compression of data results in the efficient and faster upload of data on the storage. Again we will differentiate the systems based on different compression algorithms and techniques used by different systems. The following are some important data compression algorithms for structured big data files:

- **Huffman Coding:** It is a greedy algorithm used for lossless compression of data. It makes sure that the prefix of one code should not be the prefix of another code. It works on character bits. It ensures that there should be no ambiguity while decoding the bitstream.
- **Entropy Encoding:** The father of data theory, Shannon proposed some form of entropy encoding used for lossless compression frequently. This compression technique is based on data theoretic techniques.
- **Simple Repetition:** For the n-times successive appearance of the same token sequence in a series can be replaced with a token and a count representing several appearances. A flag is used to represent whenever the repeating token appears.
- **Gzip:** GNU zip is a modern-day compression algorithm where its main function is to compress and decompress the files for faster network transfer. It reduces the size of the named file using the technique Lempel-Ziv coding. It is based on Huffman Encoding and uses an approach of LZ77 which looks at partial strings within the text.
- **Bzip2:** It is based on Burrows-Wheeler sorting text algorithm and Huffman Encoding which works on blocks that go from 100 to 900 KB. It is open to all, an open-source compression program for files, and is also free to all.

Various systems have either opted for traditional physical servers to store their big data or cloud servers to make the system cost-effective. So, we will also be differentiating the storage option used by different systems to draw a better system out of all. The following are some important storage options that can be used in a system to store data files:

- **Physical Servers:** Traditionally, Big Data is stored on physical servers and is extracted, transformed, and translated on that same server. They are generally managed, owned, and maintained by the company's staff.
- **Cloud:** It is a virtual server running in a cloud computing environment. It is accessed via the internet and can be accessed remotely and is maintained by a third party. Customers need not pay for hardware and software, rather they need to pay for resources. Some of the cloud providers are AWS, Microsoft Azure, Google Cloud Platform, IBM Cloud, Rackspace, Oracle Cloud, and Verizon Cloud.
- **Hard Disk Drives:** Major improvements are going on by the manufacturers to provide a better performance of newer 10,000 rpm 2.5-inch Hard Disk Drives than older, 15,000 rpm 3-inch devices. These advancements include heat-assisted magnetic recording which boots up the storage capacity of the device. These better Hard Disk Drives are growing rapidly and are providing a better environment to store Big Data.
- **Federated Cloud System:** Federated cloud systems are a combination of pre-existing or newly generated internal or external clouds to fulfill business needs. In this combination of several clouds, clouds may perform different actions or common action.

The efficient storage of big data is a major issue in most organizations and that's why many researchers have tried to deal with many different ways. A discussion of a few works is done here:
Jahanara et.al.[5] proposed more secure big data storage protocols. The writers implemented this using an access control model (using parameters) and honeypot (for invalid users) in a cloud computing environment. They concluded that it is needed to change faith and admission control procedure in a cloud location for big data processing. This system was suitable for analyzing, storing, and retrieval of big data in a cloud environment.
Krish et al. [6] aimed to enhance the overall I/O throughput of Big Data storage using Solid-state drives (SSD's) and for that, they designed and implemented a dynamic data management system for big data processing. The system used a map-reduce function for data processing and SSD for storage. They divided SSD into two tiers and kept a faster tier as a cache for frequently accessed data. This system, as a result, shows only 5% overhead in

performance and offers inexpensive and efficient storage management.

Hongbo et al. [7] proposed a system that reduces the movement of the data and releases intense I/O performance congestion. They came up with a combination of two data compression algorithms that effectively selects and switches to accomplish the most optimum input-output results. They experimented with a real-time application that was conducted on a cluster machine with 1280 cores and each core was comprised of 80-nodes. They come up with the result that to affect the decision for compression, processors available and the compression ratio are the two most important factors.

Eugen et al. [8] aimed to evaluate the performance and energy friendliness of various Big data tools and applications in the cloud environment which use Hadoop map-reduce. They conducted this evaluation on physical as well as on virtual clusters via different configurations. They concluded that although Hadoop is popularly used in a virtual cloud environment, we are still unable to find its correct configuration.

Mingchen et. al. [23] proposes a systematic approach to identify trends and patterns with big data. Here Big Data Analytics is applied to criminal data. Criminal data were collected and analyzed. They used the prophet model and neural network model for identifying the trend. In conclusion, they found that the prophet model works better than the neural network model. They haven't mentioned any compression technique that they used in their research.

Lu et. al. [24] aimed to explore duplicate detection in a news article. They proposed a tool NDFinder using the hash technique to detect article duplication. They checked 33,244 news articles and detects 2150 duplicate articles. Their precision reached 97%. They matched the hash values of different articles and report those having similar values. They used this research in finding plagiarism in various articles. They also found that the 3 fields with the highest

proportion of plagiarism are sports news, technology news, and military news. Moustafa et.al.[13] proposed an algorithm to minimize bandwidth cost as big data requires a lot of resources for computation and high bandwidth to allow data transfer. They proposed that a federated cloud system provides a cost-effective solution for analyzing, storing, and computing big data. This algorithm also proposed a way to minimize electricity costs for analyzing and storing big data. They concluded that their proposed algorithms perform better than the existing approach when the application is large. They proposed an algorithm that minimizes the cost of energy and bandwidth utilization to a certain extent.

Carlos et. al. [26] proposed a comparative study based on obesity based on electronic health records. They collected 20,706,947 records from different hospitals. While loading a huge amount of data they found that MongoDB shows better performance. They also found out that MySQL took double the time of MongoDB for database loading. For data query and retrieval, MySQL was found out to be more efficient. So, they concluded that MongoDB is better for database loading and MySQL is better for data query and data retrieval.

Containers reduce the downtime in the backend and provide a better service to the clients. It can be quoted from another research paper as Avanish et. al. [29] proved in their research paper that Docker containers present better results to their benchmark testing than clusters made on Virtual Machines. They also stated that cluster on containers shows better efficiency than those on Virtual Machines.

Based on the above-mentioned study a conclusive comparison has been derived which depicts how existing systems differ from each other based on our differentiating parameters. Existing systems can be categorized into the following two categories:

1. Systems with compression (in Table 1)
2. Systems without compression (in Table 2)

Findings from systems with compression:

- System 1 , there is a requirement for RDF-explicit capacity procedures and calculations for building productive and superior RDF stores.
- System 2 performs parallel compression on different processors using different compression techniques. Discovered that the proportion after compression and accessible processor are the most significant elements to affect the compression choice.

Findings from systems without compression:

- System 1 processes Big Data using Deep Learning, Neural network, and Prophet model and discovered that the Deep learning model and Prophet model and have better precision than the neural network model.
- System 2 implemented a tool NDFinder which detects duplicated data with a precision of 97%. From 33,240 records, they detected 2,150 positive duplicate records.
- System 3 reduced operation cost coming out because of big data applications getting deployed on a federated cloud.
- System 4 compares Spark and Hadoop and found out that Spark is a faster tool for Big Data processing as compared to Hadoop.
- System 5 analyzed the current technologies for big data storage and processing and came out with the result that MongoDB is better for database loading and MySQL is better for data query and data retrieval

Table 1
Comparative study of systems with compression

System No.	Paper Reference	Data Processing	Data Compression	Storage
1	Yuan et al. (2019) [28]	Hashing & Indexing	Huffman Encoding	Physical Servers
2	HongboZou et al. (2014) [7]	Not mentioned	bzip2 and gzip	Not mentioned

Table 2
Comparative study of systems without compression

System No.	Paper Reference	Data Processing	Storage
1.	Feng, Mingchen, et al. (2019) [23]	Prophet method and neural network	Cloud Storage
2.	Lu et al. (2019) [24]	NDfinder, a hashing based tool	Physical Servers
3.	MoustafaNajm et al. (2019) [13]	Virtual Machines Federated	Cloud System
4.	Cheng et al. (2019) [25]	Spark and Hadoop	Cloud Storage
5.	Carlos et al. (2019) [26]	MongoDB and MySQL	Cloud Storage
6.	Avanish et al. (2018) [29]	Docker Containers vs Hadoop	Physical Storage
7.	Pandey et al. (2016) [27]	NoSQL	Cloud Storage
8.	Krish et. al (2016)[6]	Hadoop/MapReduce and spark	Hard Disk Drive
9.	Feller, Eugen et al. (2015) [8]	Hadoop/MapReduce	Physical Servers
10.	Jahanara, et al. [5]	Hadoop/ MapReduce	Cloud Storage

- In system 6, Docker containers perform way faster than the Hadoop framework on containers.

- System 7 states that NoSQL gives a faster result than SQL

- System 8 shows Hadoop/ Map Reduce requires a higher specified environment, making the process almost impossible for low specific systems. Though parallel processing gives a better performance

- System 8 depicts that data locality is an important factor in energy efficiency.

- System 9 proposed a model to secure Big Data while storing it in the cloud. Honeypot is used as a trap to catch a thief or hacker and unauthorized user.

3. Research Question & Hypothesis

After summarizing the works depicted in Table 1 and Table 2 following gaps are identified out:

- Existing systems tried to reduce the size by compression but the results were not worthy of the time spent on it. Because compression of such a huge amount of data requires a lot of time and results, we got is not enough fast to compensate for that time.

- The transmission of big data involves a large number of bits and a lot of time is consumed. So, during transmission, there is a higher probability of data loss orbit corruption.

As the size of data is increasing day by day, servers are getting exhausted, requiring the installation of high-cost hardware. That's why there is a need to store it in such a way that it utilizes less space and should be cost-effective. The main objective is to reduce the storage size of big data and to provide an approach to keep on working with the existing physical data storage even with an increased number of users and files. The system aims to reduce data redundancy and provides a cost-effective and flexible environment.

Expected output and outcome of the solution:

A storage management system that provides Space-efficient storage of big data, reduction of data redundancy, and a cost-effective and flexible way to access data Based on the facts mentioned in the previous section our research rotates around the question that can such a system be built using different tools and algorithms which can provide efficient but faster storage of big data. This paper can give an idea of how our research towards our goal progresses. We hypothesize that different techniques can be used and combined to create a better system than the existing systems.

Design of proposed solution:

The aim is to reduce the storage space for storing big data. As the data sets are huge and also contain a lot of redundant data, first we will try to remove the redundant data from the original data set. After removing redundant data our data set would be more accurate and precise. After removing redundant data, we will compress our

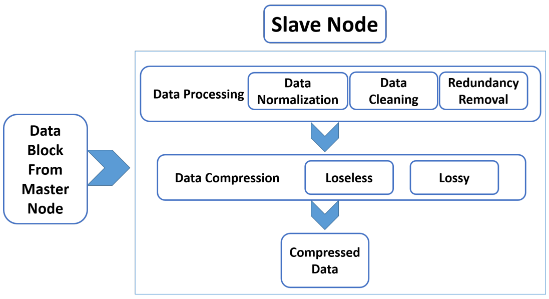

Figure 2: Proposed Architecture.

data using any good data compression technique. Then data will be stored on any desirable cloud server. We propose a system that can be produced with three modules as depicted in Fig2 and depicted below:

3.1. Data Processing:

Big data processing is a process of handling a large volume of data. Data processing is the act of manipulation of data by computer. Its main purpose is to find valuable information from raw data. Data processing can be of three types: manual data processing, batch data processing, and real-time data processing. We have the following tools and techniques for Data processing:

- Apache Giraph
- Apache Hive
- Apache Spark
- Hadoop MapReduce
- PySpark
- Prophet Method
- Neural Network Method
- MongoDB
- NoSQL
- Hashing

3.2. Data Compression:

Data Compression is the way towards diminishing the information required for capacity or transmission. It includes changing, encoding, and changing over piece structures so it consumes less space. A typical compression procedure eliminates and replaces tedious pieces and images to decrease size. There can be two types of compressions one is losing and the other is lossless. We have the following tools and techniques for Data Compression:

- Huffman Encoding
- Entropy Encoding
- Simple Repetition
- Bzip2
- Gzip

3.3. Data Storage:

Data storage refers to storing information on a storage medium. There are many storage devices available in which we can store data. Some of the storage devices are magnetic tape, disks like floppy, and ROMs, or we can store them in Cloud.

Benchmark System: The following two works can serve as the benchmark for any such projects:

- The system discussed by Haoyu et al. in their work "CloST: A Hadoop-based Storage System for Big Spatio-Temporal Data Analytics" [9]. They tried to reduce the storage space and query processing time. They proposed a three-level hierarchal partitioning approach for parallel processing for all the objects at different nodes. They used Hadoop Map-Reduce for data processing and column level gzip for data compression.
- The one discussed by Avanish et al in their work "Comparative Study of Hadoop over Containers and Hadoop Over Virtual Machine". They tried to attain a faster parallel processing environment which is developed using Docker Containers rather than the Hadoop framework on containers.

4. Conclusion

A review has been done in this paper, for data processing and compression and mentioned that each technique has its benefit for a specific type of data set. This effort is made to explore the maximum possible, important techniques coming from all the existing techniques used for the purpose. Also, we have proposed an architecture that can achieve the initial objective of reducing storage space. To achieve this, the architecture first performs analytics on the dataset, and then its size is reduced to an extent making it easy to store. After analysis and compression, the resultant dataset is stored in a cloud environment to provide better scalability and accessibility. Many types of research have been conducted already up to a scope to resolve the issue of efficient big data storage but some more persuasive steps are still essential to be taken.

References

[1] Li, Yang, and Yike Guo. ”Wiki-health: from quantified self to self-understanding.” Future Generation Computer Systems 56 (2016): 333-359.

[2] Bernard Marr. ”How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read” Forbes, 21 May 2018,

[3] Yang, Chaowei, Yan Xu, and Douglas Nebert. ”Redefining the possibility of digital Earth and geosciences with spatial cloud computing.” International Journal of Digital Earth 6.4 (2013): 297-312.

[4] Huang, Qunying, et al. ”Cloud computing for geosciences: deployment of GEOSS clearinghouse on Amazon’s EC2.” Proceedings of the ACM SIGSPATIAL international workshop on high performance and distributed geographic information systems. 2010.Sharma, Sushil, et al. ”Image Steganography using Two’s Complement.” International Journal of Computer Applications 145.10 (2016): 39-41.

[5] Akhtar, Jahanara, et al. ”Big Data Security with Access Control Model and Honeypot in Cloud Computing.” International Journal of Computer Applications 975: 8887.

[6] Krish, K. R., et al. ”On efficient hierarchical storage for big data processing.” 2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid). IEEE, 2016.

[7] Zou, Hongbo, et al. ”Improving I/O performance with adaptive data compression for big data applications.” 2014 IEEE International Parallel & Distributed Processing Symposium Workshops. IEEE, 2014.

[8] Feller, Eugen, Lavanya Ramakrishnan, and Christine Morin. ”Performance and energy efficiency of big data applications in cloud environments: A Hadoop case study.” Journal of Parallel and Distributed Computing 79 (2015): 80-89.

[9] Tan, Haoyu, Wuman Luo, and Lionel M. Ni. ”Clost: a hadoop-based storage system for big spatio-temporal data analytics.” Proceedings of the 21st ACM international conference on Information and knowledge management. 2012.

[10] Prasad, Bakshi Rohit, and Sonali Agarwal. ”Comparative study of big data computing and storage tools: a review.” International Journal of Database Theory and Application 9.1 (2016): 45-66.

[11] Schneider, Robert D. ”Hadoop for dummies.” John Willey & sons (2012).

[12] Prasetyo, Bayu, et al. ”A review: evolution of big data in developing country.” Bulletin of Social Informatics Theory and Application 3.1 (2019): 30-37.

[13] Najm, Moustafa, and Venkatesh Tamarapalli. ”Cost-efficient Deployment of Big Data Applications in Federated Cloud Systems.” 2019 11th International Conference on Communication Systems & Networks (COMSNETS). IEEE, 2019.

[14] Chatterjee, Amlan, Rushabh Jitendrakumar Shah, and Khondker S. Hasan. ”Efficient Data Compression for IoT Devices using Huffman Coding Based Techniques.” 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.

[15] Yin, Chao, et al. ”Robot: An efficient model for big data storage systems based on erasure coding.” 2013 IEEE International Conference on Big Data. IEEE, 2013.

[16] Yang, Chaowei, et al. ”Big Data and cloud computing: innovation opportunities and challenges.” International Journal of Digital Earth 10.1 (2017): 13-53.

[17] Jagadish, Hosagrahar V., et al. ”Big data and its technical challenges.” Communications of the ACM 57.7 (2014): 86-94.

[18] Bryant, Randal, Randy H. Katz, and Edward D. Lazowska. ”Big-data computing: creating revolutionary breakthroughs in commerce, science and society.” (2008).

[19] Agrawal, Divyakant, et al. ”Challenges and opportunities with Big Data 2011-1.” (2011).

[20] Xin, Luna Dong. ”Big Data Integration (Synthesis Lectures on Data Management).” (2015).

[21] Khan, Nawsher, et al. ”Big data: survey, technologies, opportunities, and challenges.” The scientific world journal 2014 (2014).

[22] Kodituwakku, S. R., and U. S. Amarasinghe. ”Comparison of lossless data compression algorithms for text data.” Indian journal of computer science and engineering 1.4 (2010): 416-425.

[23] Feng, Mingchen, et al. ”Big data analytics and mining for effective visualization and trends forecasting of crime data.” IEEE Access 7 (2019): 106111-106123.

[24] Lu, Lu, and Pengcheng Wang. ”Duplication Detection in News Articles Based on Big Data.” 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). IEEE, 2019.

[25] Cheng, Yan, Qiang Zhang, and Ziming Ye. ”Research on the Application of Agricultural Big Data Processing with Hadoop and Spark.” 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE, 2019.

[26] Martinez-Millana, Carlos, et al. ”Comparing data base engines for building big data analytics in obesity detection.” 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2019.

[27] Pandey, Manish Kumar, and Karthikeyan Subbiah. ”A novel storage architecture for facilitating efficient analytics of health informatics Big Data in cloud.” 2016 IEEE International Conference on Computer and Information Technology (CIT). IEEE, 2016.

[28] Yuan, Pingpeng, et al. ”Big RDF Data Storage, Computation, and Analysis: A Strawman’s Arguments.” 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2019.

[29] Singh, Avanish, et al. ”Comparitive Study of Hadoop over Containers and Hadoop Over Virtual Machine.” International Journal of Applied Engineering Research 13.6 (2018): 4373-4378.

[30] Sayood, Khalid. Introduction to data compression. Morgan Kaufmann, 2017.